SSR-NET: Spatial-Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion

Xueting Zhang, Wei Huang, *Student Member, IEEE,* Qi Wang, *Senior Member, IEEE,* and Xuelong Li, *Fellow, IEEE*

***Abstract*—The fusion of a low spatial resolution hyperspectral image (LR-HSI) with its corresponding high spatial resolution multispectral image (HR-MSI) to reconstruct a high spatial resolution hyperspectral image (HR-HSI) has been a significant subject in recent years. Nevertheless, it is still difficult to achieve the cross-mode information fusion of spatial mode and spectral mode when reconstructing HR-HSI for the existing methods. In this paper, based on convolutional neural network (CNN), an interpretable Spatial-Spectral Reconstruction Network (SSR-NET) is proposed for more efficient hyperspectral and multispectral image fusion. More specifically, the proposed SSR-NET is a physical straight-forward model which consists of three components: (1) Cross-Mode Message Inserting (CMMI). This operation can produce the preliminary fused HR-HSI, preserving the most valuable information of LR-HSI and HR-MSI. (2) Spatial Reconstruction Network (SpatRN). The SpatRN concentrates on reconstructing the lost spatial information of LR-HSI with the guidance of Spatial Edge Loss ($\mathcal{L}_{spat}$). (3) Spectral Reconstruction Network (SpecRN). The SpecRN pays attention to reconstruct the lost spectral information of HR-MSI under the constraint of Spatial Edge Loss ($\mathcal{L}_{spec}$). Comparative experiments are conducted on six HSI datasets of Urban, Pavia University (PU), Pavia Center (PC), Botswana, Indian Pines (IP) and Washington DC Mall (WDCM), and the proposed SSR-NET achieves the superior or competitive results in comparison with seven sate-of-the-art methods. The code of SSR-NET is available at https://github.com/hw2hwei/SSRNET.**

***Index Terms*—Hyperspectral image (HSI), multispectral image (MSI), image fusion, Cross-Mode Message Inserting, convolutional neural network (CNN), Spatial-Spectral Reconstruction Network (SSR-NET).**

I. INTRODUCTION

HYPERSPECTRAL imaging is a technology where images of hundreds of narrow spectral bands with different wavelengths can be obtained. Since hyperspectral images (HSI) have a high sepectral coverage which can accurately identify the materials and objects on the ground, there are wide applications of HSI in the fields of image classification [1]–[3], object detection [4], band selection [5]–[8], change detection [9]–[11] and so on. However, long exposures of hyperspectral systems are necessary for enough signal-to-noise ratio (SNR), which leads to the low spatial resolution of hyperspectral images (LR-HSI). In contrast, multispectral systems can acquire high spatial resolution of multispectral images (HR-MSI). Thus, it is meaningful to reconstruct high spatial resolution hyperspectral images (HR-HSI) with LR-HSI and HR-MSI, which is called as hyperspectral and multispectral image fusion .

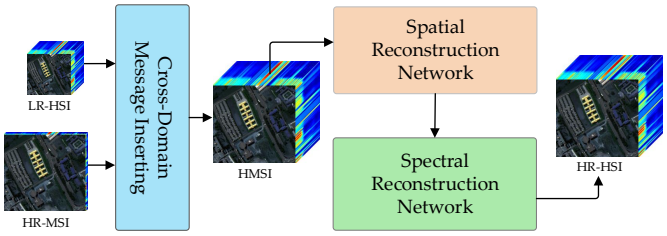


Fig. 1: Reconstructing HR-HSI from the corresponding HR-MSI and LR-HSI. Firstly, the messages of HR-MSI and LR-HSI are passed across spatial and spectral modes. Then in the orange stage, the Spatial Reconstruction Network aims to reconstruct the lost spatial information of LR-HSI. Finally in the green stage, the Spectral Reconstruction Network aims to reconstruct the lost spectral information based on the HSI which is reconstructed in spatial mode, and generates the estimated HR-HSI.

In recent years, a number of researches have been done in the field of hyperspectral and multispectral image fusion [12]–[17], which can be roughly divided into two categories, traditional methods and deep learning methods. In traditional methods, there are some different approaches including matrix factorization-based methods, Bayesian-based methods and tensor-based methods. Although these methods have achieved excellent performance on LR-HSI and HR-MSI fusion, it is still challenging to efficiently pass messages across spatial and spectral modes, which is crucial for improving the fusion quality.

Compared with traditional methods, methods based on deep learning, especially convolutional neural network (CNN) [18]–[20], have shown superior performance owing to their powerful feature-extraction ability. Hence, based on CNN, as shown in Fig. 1, a Spatial-Spectral Reconstruction Network (SSR-NET) is proposed for LR-HSI and HR-MSI fusion for the first time in this paper. Different from the previous work, the proposed method focuses on passing valuable messages across spatial and spectral modes in an efficient and interpretable manner. More concretely, the proposed SSR-NET is a three-

This work was supported by the National Key R&D Program of China under Grant 2018YFB1107403, National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, and 61871470.

X. Zhang, W. Huang, Q. Wang and X. Li are with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zxt@mail.nwpu.edu.cn; hw2hwei@gmail.com; crabwq@gmail.com; li@nwpu.edu.cn).

X. Li is the corresponding author.

stage network model made up of three modules: (1) Cross-Mode Message Inserting (CMMI). (2) Spatial Reconstruction Network (SpatRN) with the Spatial Edge Loss optimized by $\mathcal{L}_{spat}$. (3) Spectral Reconstruction Network (SpecRN) with the Spectral Edge Loss optimized by $\mathcal{L}_{spec}$.

Firstly, as the input of SSR-NET, the LR-HSI and HR-MSI are sent into the CMMI, which maintains the valuable spectral information of LR-HSI and spatial information of HR-MSI, and then fuses them into a hyper-multiple spectral image (HMSI) which has the same size as the reference HR-HSI (*i. e.*, the ground-truth HR-HSI). Secondly, the SpatRN is designed to reconstruct the lost spatial information of HMSI. However, CNN is a black-box model and the learned features are lack of enough explaination. To assign the physical meaning to Spatial Reconstruction Network, the spatial edge loss of $\mathcal{L}_{spat}$ is presented to constrain SpatRN to focus on spatial reconstruction. Finally, the SpecRN is designed to reconstruct the lost spectral information from the reconstructed HMSI in spatial mode. Similarly, the spectral edge loss of $\mathcal{L}_{spec}$ is designed to make SpecRN pay attention to spectral restoration.

In summary, the main contributions of this paper conclude the following aspects:

1) Based on CNN, a novel Spatial-Spectral Reconstruction Network (SSR-NET) is first proposed for more efficient hyperspectral and multispectral image fusion in this paper.
2) The proposed SSR-NET is a physical straight-forward CNN model, which is under the constraint of Spatial Edge Loss $\mathcal{L}_{spat}$ and Spectral Edge Loss $\mathcal{L}_{spec}$. $\mathcal{L}_{spat}$ and $\mathcal{L}_{spec}$ are specially designed for spatial and spectral reconstruction.
3) Compared with seven state-of-the-art approaches, the proposed SSR-NET achieves the best results on five HSI datasets of Urban, Pavia University (PU), Pavia Center (PC), Botswana and Indian Pines (IP), and the competitive results on the dataset of Washington DC Mall (WDCM). Such experimental results demonstrate the effectiveness and superiority of the proposed SSR-NET in LR-HSI and HR-MSI fusion.

The remainder of this paper is organized as follows: Section II gives an introduction of the related works of LR-HSI and HR-MSI fusion. In Section III, we describe the proposed SSR-NET model in detail. The experimental results on four datasets are analyzed in Section IV. In the end, conclusions are provided in Section V.

II. Related Work

In this section, some classical methods of HSI and MSI fusion will be reviewed, which roughly fall into two categories: traditional methods and deep learning methods.

A. Traditional Methods

Generally, traditional methods include the following three types of methods:

1) Matrix Factorization-Based Methods: This type of methods usually unfold the 3-D HSI, where the three dimensions respectively denote the width, height, and the number of spectral bands, into a 2-D matrix, where the two dimensions denote the flattened spatial locations and the band number. In these methods, two matrices of an endmember matrix and an abundance matrix are expected to be respectively estimated from the LR-HSI and the HR-MSI, and are used to reconstruct the corresponding HR-HSI.

For example, based on unsupervised unmixing, Yokoya et al. [12] present a coupled nonnegative matrix factorization (CNMF) method for LR-HSI and HR-MSI fusion, where the abundance matrix of high spatial resolution acquired from multispectral images and the hyperspectral endmember matrix are integrated to generate a new HR-HSI. Then Lanaras et al. [13] propose a method which concentrates on jointly the spectral unmixing problem for both two input images, where some constraints are added according to the physical properties of spectral unmixing. In [14], Dong et al. present a sparsity-based hyperspectral image super-resolution method on the basis of dictionary learning and sparse representation. After that, considering the ill-posed inverse problem, an ADMM-based CO-CNMF algorithm is proposed by Lin et al. [21] for further optimization of CNMF, where ℓ_1 norm and SSD regularizers are incorporated and added to CNMF criterion.

2) Bayesian-Based Methods: This type of methods generally utilize the appropriate prior distribution of the given images to solve the fusion problem of LR-HSI and HR-MSI.

For instance, a Bayesian sparse representation based approach is first proposed by Akhtar et al. [15]. In this work, the non-parametric Bayesian dictionary learning is utilized to learn the distributions of the scene spectrum and the proportion of them in the image, where the distributions would be utilized to compute sparse codes of the high resolution image. In [22], Wei et al. propose a fast multi-band image fusion algorithm (FUSE) which is based on a solution of a Sylvester equation. Combining the alternating direction method of multiplier and block coordinate descent, this algorithm can be easily extended to calculate Bayesian estimators of fusion.

3) Tensor Factorization-Based Methods: Different from matrix factorization-based methods, methods based on tensor factorization typically treat the HSI as a 3-D tensor, which has three modes as described above. In these methods, the HR-HSI are cut apart into some cubes, and then the similar cubes would be grouped based on the learned clusters and some certain priors.

For example, on the basis of Tucker factorization, Dian et al. first propose a novel HSI super-resolution method [16] which is called as the NLSTF. It unifies the sparse tensor factorization and the non-local means approach into one framework, and thus considering the HSI super-resolution problem from the point of the estimation of dictionaries and a sparse core tensor for each cube. Similarly, a coupled sparse tensor factorization based framework named as CSTF is proposed in [23], where the estimation of the core tensor and three dictionaries are formulated as a coupled sparse tensor decomposition of the HR-MSI and LR-HSI. In [24], taking geometric structures into account, a spatial-spectral-graph-regularized low-rank tensor

decomposition method (SSGLRTD) is presented by Zhang et al. Meanwhile, a ALM based algorithm is specially designed for better optimization of the fusion model. Following the previous work in [16], Dian et al. design a novel LTTR prior in [25] to learn the relationship among the spectral, spatial, and nonlocal modes of the nonlocal similar HR-HSI cubes. In this work, the similar HR-HSI cubes are composed into a 4-D tensor, and the ADMMs [26] algorithm is adopted to solve the optimization problem.

B. Deep Learning Methods

With the rapid develpoment of deep learning methods, especially convolutional network (CNN) [18]–[20], [27]–[29], these types of methods have become a growing trend in all kinds of hyperspectral image processing [30]–[32]. In LR-HSI and HR-MSI fusion, deep learning methods show the excellent performance. Different from traditional methods, in deep learning-based approaches, various neural networks are usually exploited to enhance the performance of image fusion in a learnable manner.

For example, Palsson et al. [33] first propose a 3-D-Convolutional Neural Network (CNN) to acquire HR-HSI from LR-HSI and HR-MSI, where PCA prior is utilized for dimensionality reduction of the fusion. After that, a deep HSI sharpening method (DHSIS) is proposed by Dian et al. [17], which learns the priors by residual learning to achieve the regularization of image fusion problem. In [34], a novel MS/HS fusion network is presented by Xie et al., which considers the generation mechanism underlying the MSI/HSI fusion data. Following that, based on CNN denoiser and subspace representation, Dian et al. [35] propose a new HSI-MSI fusion method which can be applied to different HSI datasets without retraining. By utilizing techniques of network-in-network convolutional unit, skip connection, and batch normalization, Xu et al. [36] propose a two-branch network (HAM-MFN) to gradually reconstruct HR-HSI via fusing LR-HSI and HR-MSI at different scales, where a RAP loss is designed to deal with spectral and spatial distortions. And some other related deep learning methods including TFNet [37], ResTFNet [37], SSFCNN [38], ConSSFCNN [38] and MSDCNN [39] are introduced in detail in Section IV as the comparative methods.

III. Methodology

In this section, the proposed Spatial-Spectral Reconstruction Network (SSR-NET), which is shown in Fig. 2, will be introduced in detail. Overall, the proposed SSR-NET is a physical straight-forward CNN model and it mainly consists of three modules: (1) Cross-Mode Message Inserting (CMMI). (2) Spatial Reconstruction Network (SpatRN) with Spatial Edge Loss ($\mathcal{L}_{spat}$). (3) Spectral Reconstruction Network (SpecRN) with Spectral Edge Loss ($\mathcal{L}_{spec}$).

A. Cross-Mode Message Inserting

In the proposed SSR-NET, the reference HR-HSI and the estimated HR-HSI are denoted as $\mathbf{R} \in \mathbb{R}^{H\times W\times L}$ and $\mathbf{Z} \in \mathbb{R}^{H\times W\times L}$, where H and W respectively represent the dimensions of the height and width, and L represents the number of spectral bands. Besides, its inputs are a LR-HSI denoted as $\mathbf{X} \in \mathbb{R}^{h\times w\times L}$ ($h \ll H, w \ll W$) and a HR-MSI denoted as $\mathbf{Y} \in \mathbb{R}^{H\times W\times l}$ ($l \ll L$). $\mathbf{X}$ and $\mathbf{Y}$ are respectively sampled in the spatial and spectral mode, which are obtained by:

$$\begin{cases} \mathbf{X} = Gaussian(\mathbf{Z}), \\ \mathbf{X} = Bilinear(\mathbf{X}, 1/r), \end{cases} \tag{1}$$

$$\begin{cases} \mathbf{Y}(k) = \mathbf{Z}(s_k), k \in \{1, ..., l\}, \\ s_k = (k-1) * L/(l-1), s_k \in \{s_1, ..., s_l\} \end{cases} \tag{2}$$

where $\mathbf{X}$ is spatially downsampled by bilinear operation at the ratio of r from $\mathbf{Z}$, which is blurred by Gaussian filter in advance. $\mathbf{Y}$ is sampled from $\mathbf{Z}$ at an equal interval of bands. $\mathbf{Y}(k)$ represents the k-th band of $\mathbf{Y}$, and $\{s_1, ..., s_l\}$ represent the sampled band numbers in HR-HSI.

The goal of Cross-Mode Message Inserting (CMMI) is to produce a preliminary concatenated hyper-multiple spectral image (HMSI) represented as $\mathbf{Z}_{pre} \in \mathbb{R}^{H\times W\times L}$, which takes advantages of both the spatial information of HR-MSI and the spectral information of LR-HSI preserving their relative spatial-spectral position.

By utilizing bilinear interpolation, the LR-HSI of $\mathbf{X}$ would be upsampled to the same size as the HR-MSI of $\mathbf{Y}$ in the spatial mode, which is denoted as:

$$\mathbf{X}\uparrow = Bilinear(\mathbf{X}, r), \tag{3}$$

where r is the upsampling ratio, and $\mathbf{X}\uparrow$ is the upsampled LR-HSI.

Then the HR-MSI and the upsampled LR-HSI would be preliminarily fused, which is formulated as:

$$\mathbf{Z}_{pre}(k) = \begin{cases} \mathbf{Y}(k), & \text{if } k \in \{s_1, ..., s_l\}, \\ \mathbf{X}(k)\uparrow, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathbf{Z}_{pre}$ denotes the HMSI, and $\mathbf{Z}_{pre}(k)$ is the k-th band of HMSI. Similarly, $\mathbf{Y}(k)$ and $\mathbf{X}(k)\uparrow$ represent the k-th band of HR-MSI and the upsampled LR-HSI, respectively. In this way, the concatenated HMSI can contain both spatial and spectral information of HR-MSI and LR-HSI.

To preliminarily pass the information between the modes of spatial and spectral, a convolutional layer with the kernel size set to 3×3 and spatial height and width stride set to 1 is applied in HMSI. It is denoted as:

$$\mathbf{Z}_{pre} = ReLU(Conv_{pre}(\mathbf{Z}_{pre})), \tag{5}$$

where $ReLU$ is the non-linear activation function of Rectified Linear Unit [40].

B. Spatial Reconstruction Network with Spatial Edge Loss

In order to reconstruct the spatial information from $\mathbf{Z}_{pre}$, another 3×3 convolutional layer with height/width stride set to 1 serves as the Spatial Reconstruction Network (SpatRN), which is formulated as:

$$\mathbf{Z}_{spat} = \mathbf{Z}_{pre} + Conv_{spat}(\mathbf{Z}_{pre}), \tag{6}$$

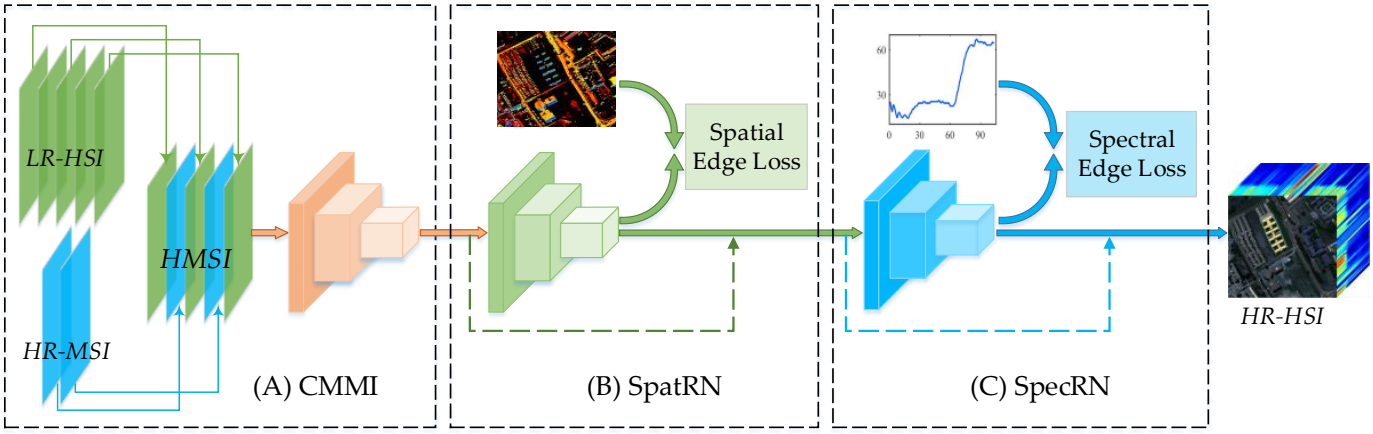


Fig. 2: The framework of the proposed SSR-NET.

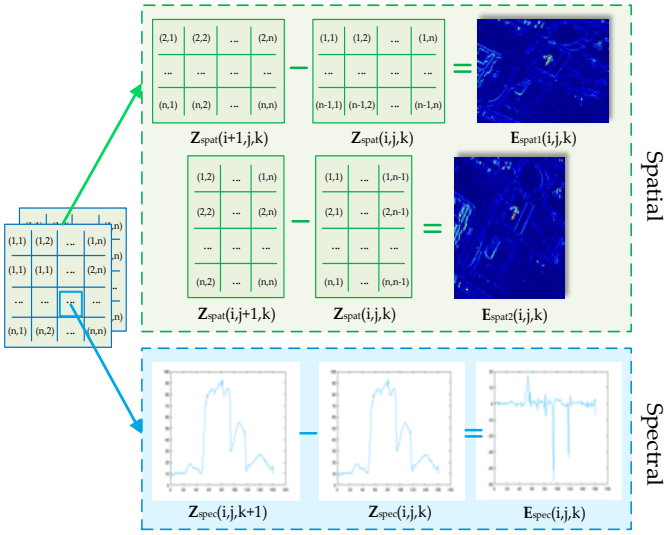


Fig. 3: The illustration of spatial and spectral edges.

TABLE I: The architecture and optimization loss of the variant models of the proposed SSR-NET on Urban dataset.

Model	Input	Architecture	Loss
Spat-CNN	LR-HSI ($\mathbf{X}$)	Conv(3×3, L, L) Conv(3×3, L, L)	 $\mathcal{L}_{fus}$
Spec-CNN	HR-MSI ($\mathbf{Y}$)	Conv(3×3, l, L) Conv(3×3, L, L)	 $\mathcal{L}_{fus}$
SpatR-NET	HMSI ($\mathbf{Z}_{pre}$)	Conv(3×3, L, L) Conv(3×3, L, L)	$\mathcal{L}_{spat}$ $\mathcal{L}_{fus}$
SpecR-NET	HMSI ($\mathbf{Z}_{pre}$)	Conv(3×3, l, L) Conv(3×3, L, L)	$\mathcal{L}_{spec}$ $\mathcal{L}_{fus}$
SSR-NET	HMSI ($\mathbf{Z}_{pre}$)	Conv(3×3, L, L) Conv(3×3, L, L) Conv(3×3, L, L)	 $\mathcal{L}_{spat}$ $\mathcal{L}_{spec}$+$\mathcal{L}_{fus}$

where $Conv_{spat}$ represents the convolutional layer. It is worth mentioning that the skip-connection operation is used to improve the model stability during the training stage.

However, due to the black-box characteristic of CNN, it is uncontrollable for the learned feature mapping. As we all know, the spatial edges of images contain the high frequency feature, which are crucial for spatial reconstruction. To make SpatRN focus on the restoration of spatial information, based on spatial edges as shown in Fig. 3, a novel Spatial Edge Loss ($\mathcal{L}_{spat}$) is proposed in this paper to constrain the output of SpatRN. $\mathcal{L}_{spat}$ is calculated by:

$$\mathbf{E}_{spat1}(i,j,k) = \mathbf{Z}_{spat}(i+1,j,k) - \mathbf{Z}_{spat}(i,j,k), \tag{7}$$

$$\mathbf{E}_{spat2}(i,j,k) = \mathbf{Z}_{spat}(i,j+1,k) - \mathbf{Z}_{spat}(i,j,k), \tag{8}$$

$$\bar{\mathbf{E}}_{spat1}(i,j,k) = \mathbf{Z}(i+1,j,k) - \mathbf{Z}(i,j,k), \tag{9}$$

$$\bar{\mathbf{E}}_{spat2}(i,j,k) = \mathbf{Z}(i,j+1,k) - \mathbf{Z}(i,j,k), \tag{10}$$

$$\mathcal{L}_{spat1} = \frac{\sum_{k=1}^{L}\sum_{i=1}^{H-1}\sum_{j=1}^{W}(\mathbf{E}_{spat1}(i,j,k) - \bar{\mathbf{E}}_{spat1}(i,j,k))^2}{2WL(H-1)}, \tag{11}$$

$$\mathcal{L}_{spat1} = \frac{\sum_{k=1}^{L}\sum_{i=1}^{H}\sum_{j=1}^{W-1}(\mathbf{E}_{spat2}(i,j,k) - \bar{\mathbf{E}}_{spat2}(i,j,k))^2}{2HL(W-1)}, \tag{12}$$

$$\mathcal{L}_{spat} = 0.5 * \mathcal{L}_{spat1} + 0.5 * \mathcal{L}_{spat2}, \tag{13}$$

where $\mathbf{E}_{spat1} \in \mathbb{R}^{(H-1)\times W\times L}$ and $\mathbf{E}_{spat2} \in \mathbb{R}^{H\times(W-1)\times L}$ are the edge maps of $\mathbf{Z}_{spat}$ along spatial height and width, respectively. Similarly, $\bar{\mathbf{E}}_{spat1} \in \mathbb{R}^{(H-1)\times W\times L}$ and $\bar{\mathbf{E}}_{spat2} \in \mathbb{R}^{H\times(W-1)\times L}$ are the edge maps of the reference HR-HSI of $\mathbf{Z}_{spat}$ along spatial height and width, respectively. $\mathcal{L}_{spat1}$ and $\mathcal{L}_{spat2}$ are the spatial height edge loss and spatial width edge loss, which are calculated between $\mathbf{E}_{spat}$ and $\bar{\mathbf{E}}_{spat}$ by the loss function of Mean Squared Error (MSE). Finally, $\mathcal{L}_{spat1}$ and $\mathcal{L}_{spat2}$ are fused into $\mathcal{L}_{spat}$.

C. Spectral Reconstruction Network with Spectral Edge Loss

After spatial reconstruction, one more convolutional layer serving as the Spectral Reconstruction Network (SpecRN) with the same architecture as SpatRN is used to further reconstruct the spectral information based on $\mathbf{Z}_{spat}$. It is formulated as:

$$\mathbf{Z}_{spec} = \mathbf{Z}_{spat} + Conv_{spec}(\mathbf{Z}_{spat}), \tag{14}$$

where $Conv_{spec}$ represents the convolutional layer. And the skip-connection operation is also applied in SpatRN.

Similar to spatial edges, as shown in Fig. 3, the spectral edges of bands contain the high frequency information that are crucial for spectral reconstruction. To make SpecRN pay attention to spectal restoration, the Spectral Edge Loss ($\mathcal{L}_{spec}$) is also proposed in this paper to constrain the output of SpecRN. $\mathcal{L}_{spec}$ is calculated by:

$$\mathbf{E}_{spec}(i,j,k) = \mathbf{Z}_{spec}(i,j,k+1) - \mathbf{Z}_{spec}(i,j,k), \quad (15)$$

$$\bar{\mathbf{E}}_{spec}(i,j,k) = \mathbf{Z}(i,j,k+1) - \mathbf{Z}(i,j,k), \quad (16)$$

$$\mathcal{L}_{spec} = \frac{\sum_{k=1}^{L}\sum_{i=1}^{H-1}\sum_{j=1}^{W}(\mathbf{E}_{spec}(i,j,k) - \bar{\mathbf{E}}_{spec}(i,j,k))^2}{2HW(L-1)}, \quad (17)$$

where $\mathbf{E}_{spec} \in \mathbb{R}^{H\times W\times(L-1)}$ and $\bar{\mathbf{E}}_{spec} \in \mathbb{R}^{H\times W\times(L-1)}$ are the edge maps of $\mathbf{Z}_{spec}$ and $\mathbf{Z}$ along the spectral mode, respectively. The spectral edge loss of $\mathcal{L}_{spec}$ is the MSE loss of $\mathbf{E}_{spec}$ and $\bar{\mathbf{E}}_{spec}$. After CMMI, SpatRN and SpecRN, $\mathbf{Z}_{spec}$ is used as the final estimated HR-HSI denoted as $\mathbf{Z}_{fus}$. It is fomulated as:

$$\mathbf{Z}_{fus} = \mathbf{Z}_{spec}, \quad (18)$$

For $\mathbf{Z}_{fus}$, it is optimized by the fusion loss denoted as $\mathcal{L}_{spec}$, which is formulated as:

$$\mathcal{L}_{fus} = \frac{\sum_{k=1}^{L}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{Z}_{fus}(i,j,k) - \bar{\mathbf{Z}}(i,j,k))^2}{2WHL}. \quad (19)$$

In the proposed SSR-NET, the overall loss denoted as $\mathcal{L}$ is the sum of $\mathcal{L}_{spat}$, $\mathcal{L}_{spec}$ and $\mathcal{L}_{fus}$:

$$\mathcal{L} = \mathcal{L}_{spat} + \mathcal{L}_{spec} + \mathcal{L}_{fus}, \quad (20)$$

D. Variations of SSR-NET

To explore the effectiveness of the components of SSR-NET including CMMI, SpatRN and SpecRN, five variant models of Spat-CNN, Spec-CNN, SpatR-NET, SpecR-NET and SSR-NET are designed for ablation experiments.

The architectures and the layer-wise loss of these five models are provided in Table I. Under the constraints of fusion loss of $\mathcal{L}_{fus}$, Spat-CNN and Spec-CNN aim to reconstruct the HR-HSI using LR-HSI and HR-MSI respectively. In contrast, the inputs of SpatR-NET, SpecR-NET and SSR-NET are all the HMSI. The SpatR-NET focuses on restoring the spatial information of the HMSI with the guidance of $\mathcal{L}_{fus}$ and the spatial edge loss of $\mathcal{L}_{spat}$, while the SpatR-NET focuses on restoring the spectral information of the HMSI with the guidance of $\mathcal{L}_{fus}$ and the spectral edge loss of $\mathcal{L}_{spec}$. SSR-NET is the combination of SpatR-NET and SpecR-NET. In this table, Conv(3×3, l, L) denotes the 3 × 3 convolutional layer, whose input and output channel number are respectively l and L with the height/width stride set to 1. Besides, each convolutional layer is followed by the non-linear activation function of $ReLU$.

IV. Experiments

In this section, sufficient experiments are conducted on six datasets to verify the effectiveness of the proposed SSR-NET.

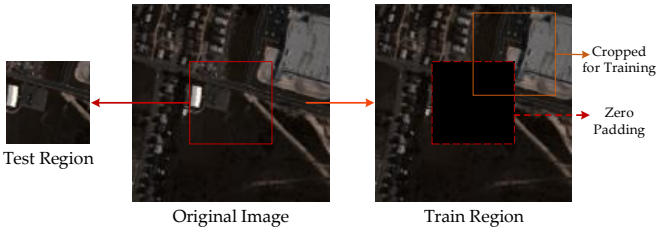


Fig. 4: An example of the train and test regions of Urban dataset. During the training stage, the test region in the dataset is padded by zeros.

First, the experimental datasets and evaluation metrics are introduced. Then the experimental settings are provided in detail. Following that, the ablation experiments of SSR-NET are performed to explore the role of its components. Moreover, the comparison experiments of SSR-NET and some state-of-the-art methods are conducted. Finally, we quantitatively analyse the interpretability of SSR-NET and make a comparison between SSR-NET and some other deep learning methods in model size and test time.

A. Datasets

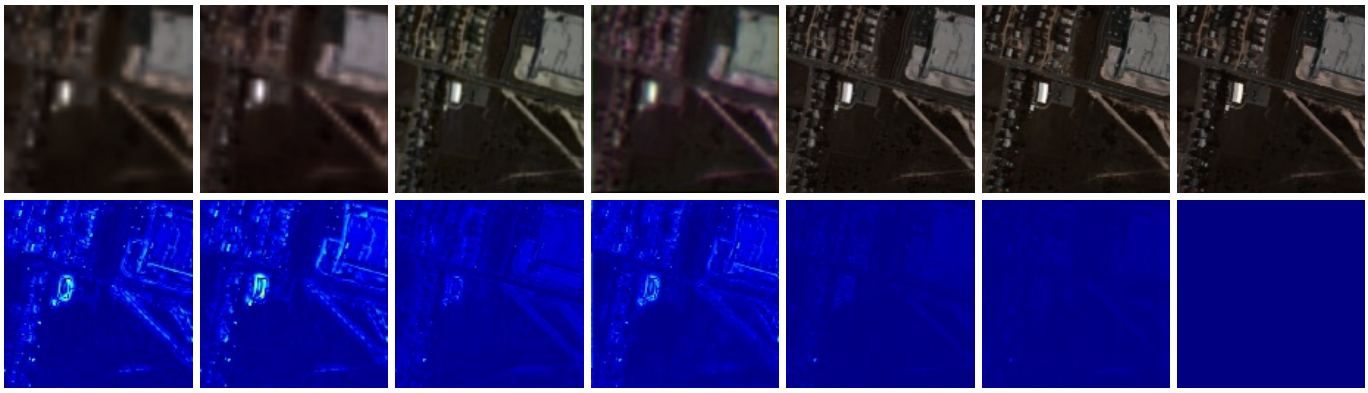
In this paper, six datasets are adopted to verify the effectiveness and generalization ability of the proposed SSR-NET, including Urban, Pavia University (PU), Pavia Center (PC), Botswana, Indian Pines (IP) and Washington DC Mall (WDCM).

1) Urban: The HYDICE Urban dataset was obtained over Copperas Cove, TX, USA, in 1995. In this dataset, there are 210 bands in total which cover the wavelengths from 0.4 to 2.5 μm with an interval of 10 nm. After removing the bands of dense water vapor and atmospheric, the remaining 162 bands are used in this paper, and the image of each band measures 307 × 307 pixels with a spatial resolution of 2 m.

2) Pavia University: The Pavia University (PU) dataset was obtained by the Reflective Optics Spectrographic Imaging System (ROSIS) sensor over Piavia University, Italy in 2003. In this dataset, there are 103 bands covering the spectral range from 0.43 to 0.86 μm with an interval of 10 nm, and the image of each band measures 610 × 340 pixels with a spatial resolution of 1.3 m.

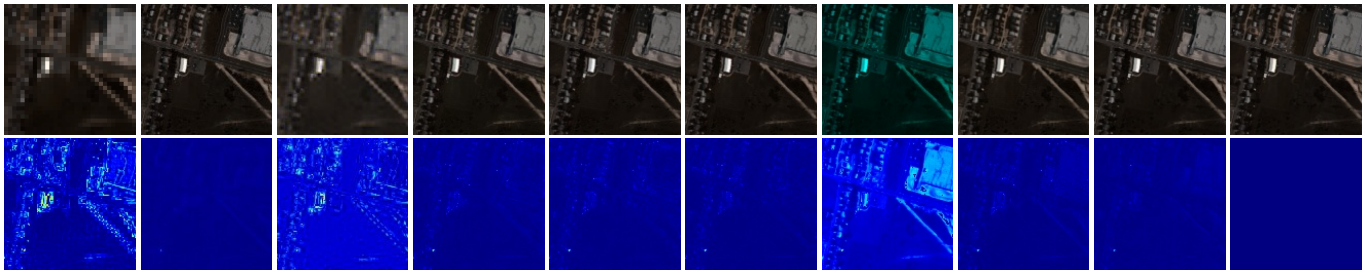
3) Pavia Center: The Pavia Center (PC) dataset was obtained by the same hyperion sensor of PU dataset with the same spatial resolution as PU dataset. However, it has one more band than PU dataset with the total number of bands as 103. The image of each band measures 1096 × 1096 pixels, which is much larger than PU dataset.

4) Botswana: The Botswana dataset was obtained by the hyperion sensor of the NASA EO-1 satellite over the Okavango Delta, Botswana in 2001-2004. In Botswana, there are 242 bands in total which cover the spectral range from 0.4 to 2.5 μm with an interval of 10 nm. After removing the uncalibrated and noisy bands of water absorption features, the rest 145 bands are remained, and the image of each band measures 1476 × 256 pixels with a spatial resolution of 30 m.



(a) LR-HSI ↑ (b) Spat-CNN (c) Spec-CNN (d) SpatR-NET (e) SpecR-NET (f) SSR-NET (g) GT

Fig. 5: The fusion results of Urban based on different model variations, where 'GT' refers to the ground-truth image. The first row shows the R-G-B images (26-11-1 bands) of the estimated HR-HSIs, and the second row shows the difference images between the estimated R-G-B image and the reference R-G-B image (*i. e.*, the ground-truth R-G-B image), which are processed by pseudocolor technique.



(a) LR-HSI ↑ (b) CNMF (c) LTTR (d) MSDCNN (e) TFNet (f) ResTFNet (g) SSFCNN (h) ConSSF (i) SSR-NET (j) GT

Fig. 6: The fusion results of different methods on Urban dataset, where 'ConSSF' and 'GT' reprensent the ConSSFCNN and the ground-truth image. The first row shows the R-G-B images (26-11-1 bands) of the estimated HR-HSIs, and the second row shows the difference images between the estimated and the reference R-G-B images, which are processed by pseudocolor technique.

5) Indian Pines: The Indian Pines (IP) dataset was obtained by AVIRIS sensor over the Indian Pines test site, Indiana. In this dataset, there are 224 bands covering the spectral range from 0.4 to 2.5 μm. After removing the bands covering the region of water absorption, there are 200 bands remained, and the image of each band measures 145 × 145 pixels.

6) Washington DC Mall: The Washington DC Mall (WDCM) dataset was obtained by the sensor of hyperspectral digital imagery collection experiment (HYDICE) over the National Mall in Washington, DC in 1995. In WDCM, there are 210 bands in total which cover the wavelengths from 0.4 to 2.5 μm. After removing the bands of water vapor absorption, the remaining 191 bands are used in this paper, and the image of each band measures 1280 × 307 pixels with a spatial resolution of 2.5 m.

B. Evaluation Metrics

Four widely used metrics are employed to evaluate the performance of the proposed SSR-NET and the comparison methods. Here, $\mathbf{R}_k(i,j)$ and $\mathbf{Z}_k(i,j)$ denote the element value at spatial location (i,j) in band k of the reference HR-HSI and the estimated HR-HSI.

1) Root Mean Squared Error (RMSE): It can be used to measure the difference between $\mathbf{R}$ and $\mathbf{Z}$ which is defined as:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{L}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{R}_k(i,j)-\mathbf{Z}_k(i,j))^2}{HWL}}, \tag{21}$$

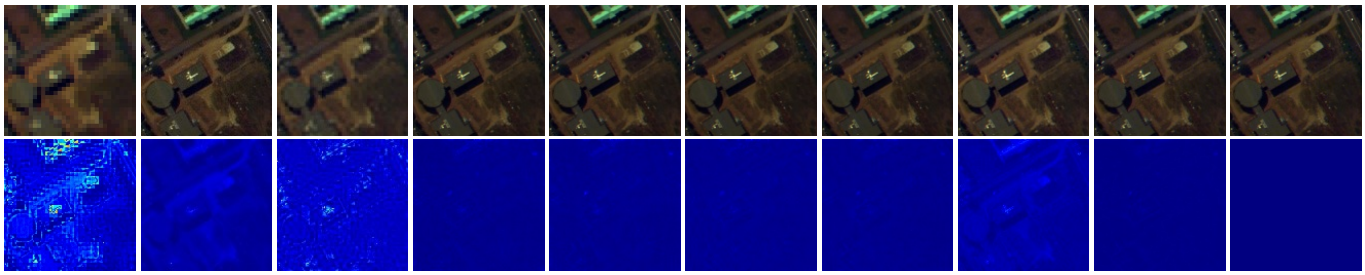
where the smaller the RMSE is, the better the performance is.

2) Peak Signal-to-Noise Ratio (PSNR): The PSNR can evaluate the spatial quality of the reconstructed HR-HSI in unit of band. The PSNR of k-th band is defined as:

$$PSNR = 10\log_{10}\left(\frac{\max(\mathbf{R}_k)^2}{\frac{1}{HW}\|\mathbf{R}_k-\mathbf{Z}_k\|_2^2}\right), \tag{22}$$

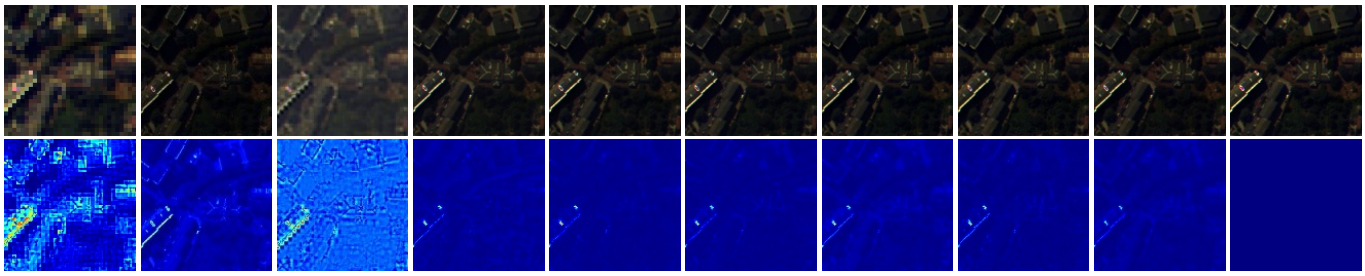
where $\mathbf{R}_k$ and $\mathbf{Z}_k$ respectively represent reference image and the estimated image of the k-th band. And $\|\cdot\|_2$ refers to the 2-norm. The final PSNR is the average of the PSNRs of all bands. The higher the PSNR is, the better the performance is.

3) Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS): The ERGAS [41] is specially designed for assess-



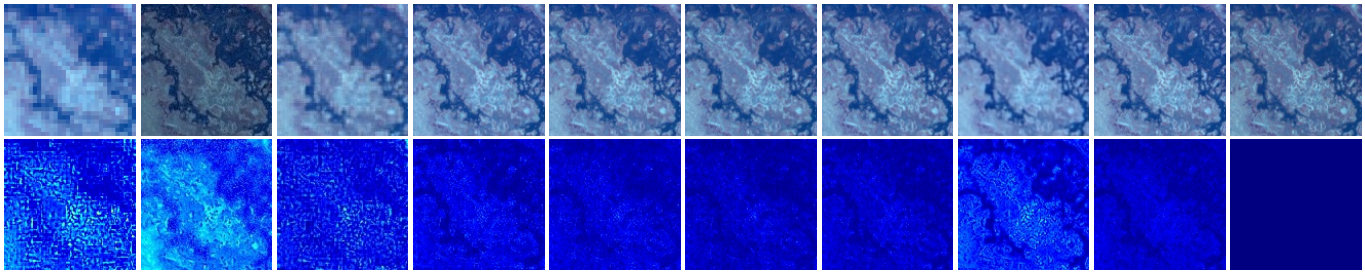
(a) LR-HSI ↑ (b) CNMF (c) LTTR (d) MSDCNN (e) TFNet (f) ResTFNet (g) SSFCNN (h) ConSSF (i) SSR-NET (j) GT

Fig. 7: The fusion results of different methods on Pavia University dataset, where 'ConSSF' and 'GT' respectively reprensent the ConSSFCNN and the ground-truth image. The first row shows the R-G-B images (67-29-1 bands) of the estimated HR-HSIs, and the second row shows the difference images between the estimated and the reference R-G-B images, which are processed by pseudocolor technique.



(a) LR-HSI ↑ (b) CNMF (c) LTTR (d) MSDCNN (e) TFNet (f) ResTFNet (g) SSFCNN (h) ConSSF (i) SSR-NET (j) GT

Fig. 8: The fusion results of different methods on Pavia Center dataset, where 'ConSSF' and 'GT' reprensent the ConSSFCNN and the ground-truth image. The first row shows the R-G-B images (67-29-1 bands) of the estimated HR-HSIs, and the second row shows the difference images between the estimated and the reference R-G-B images, which are processed by pseudocolor technique.



(a) LR-HSI ↑ (b) CNMF (c) LTTR (d) MSDCNN (e) TFNet (f) ResTFNet (g) SSFCNN (h) ConSSF (i) SSR-NET (j) GT

Fig. 9: The fusion results of different methods on Botswana dataset, where 'ConSSF' and 'GT' reprensent the ConSSFCNN and the ground-truth image. The first row shows the R-G-B images (48-15-4 bands) of the estimated HR-HSIs, and the second row shows the difference images between the estimated and the reference R-G-B images, which are processed by pseudocolor technique.

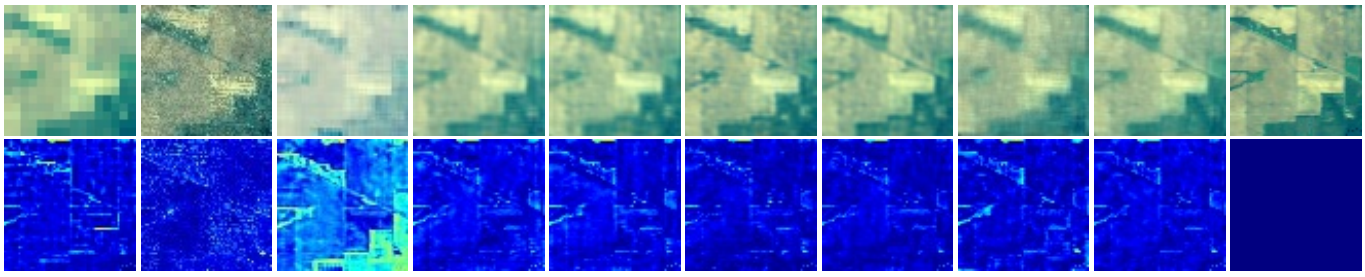
ing the quality of high resolution synthesised images, which measures the global statistical quality of the estimated HR-HSI. It is defined as:

$$ERGAS = \frac{100}{r}\sqrt{\frac{1}{L}\sum_{k=1}^{L}\frac{\|\mathbf{R}_k - \mathbf{Z}_k\|_2^2}{\mu^2(\mathbf{R}_k)}}, \tag{23}$$

where r refers to the ratio of the spatial down-sampling ratio from HR-HSI to LR-HSI. And $\mu(\mathbf{R}_k)$ denotes the mean value of the reference image of the k-th band. The smaller the ERGAS is, the better the performance is.

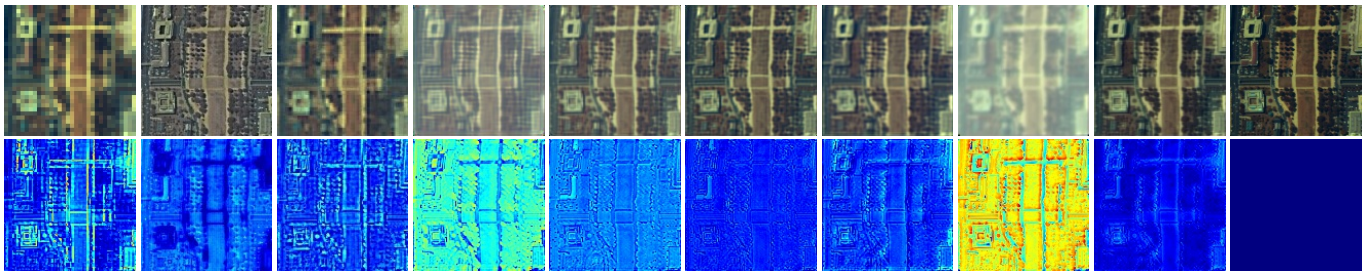
4) Spectral Angle Mapper (SAM): The SAM [42] is generally utilized to evaluate the spectral information preservation degree at each pixel, which is defined as:

$$SAM = \arccos\left(\frac{\langle \mathbf{R}(i,j), \mathbf{Z}(i,j)\rangle}{\|\mathbf{R}(i,j)\|_2 \|\mathbf{Z}(i,j)\|_2}\right), \tag{24}$$



(a) LR-HSI ↑ (b) CNMF (c) LTTR (d) MSDCNN (e) TFNet (f) ResTFNet (g) SSFCNN (h) ConSSF (i) SSR-NET (j) GT

Fig. 10: The fusion results of different methods on Indian Pines dataset, where 'ConSSF' and 'GT' reprensent the ConSSFCNN and the ground-truth image. The first row shows the R-G-B images (29-15-4 bands) of the estimated HR-HSIs, and the second row shows the difference images between the estimated and the reference R-G-B images, which are processed by pseudocolor technique.



(a) LR-HSI ↑ (b) CNMF (c) LTTR (d) MSDCNN (e) TFNet (f) ResTFNet (g) SSFCNN (h) ConSSF (i) SSR-NET (j) GT

Fig. 11: The fusion results of different methods on Washington DC Mall dataset, where 'ConSSF' and 'GT' reprensent the ConSSFCNN and the ground-truth image. The first row shows the R-G-B images (55-35-11 bands) of the estimated HR-HSIs, and the second row shows the difference images between the estimated and the reference R-G-B images, which are processed by pseudocolor technique.

where $\mathbf{R}(i,j)$ and $\mathbf{Z}(i,j)$ respectively denote the spectral vector of the reference and the estimated HR-HSI at the pixel position of (i,j). Besides, $\langle \mathbf{R}(i,j), \mathbf{Z}(i,j) \rangle$ refers to the inner product of $\mathbf{R}(i,j)$ and $\mathbf{Z}(i,j)$. The overall SAM is the average of the SAMs of all pixels. The lower the SAM is, the better the performance is.

C. Experimental Settings

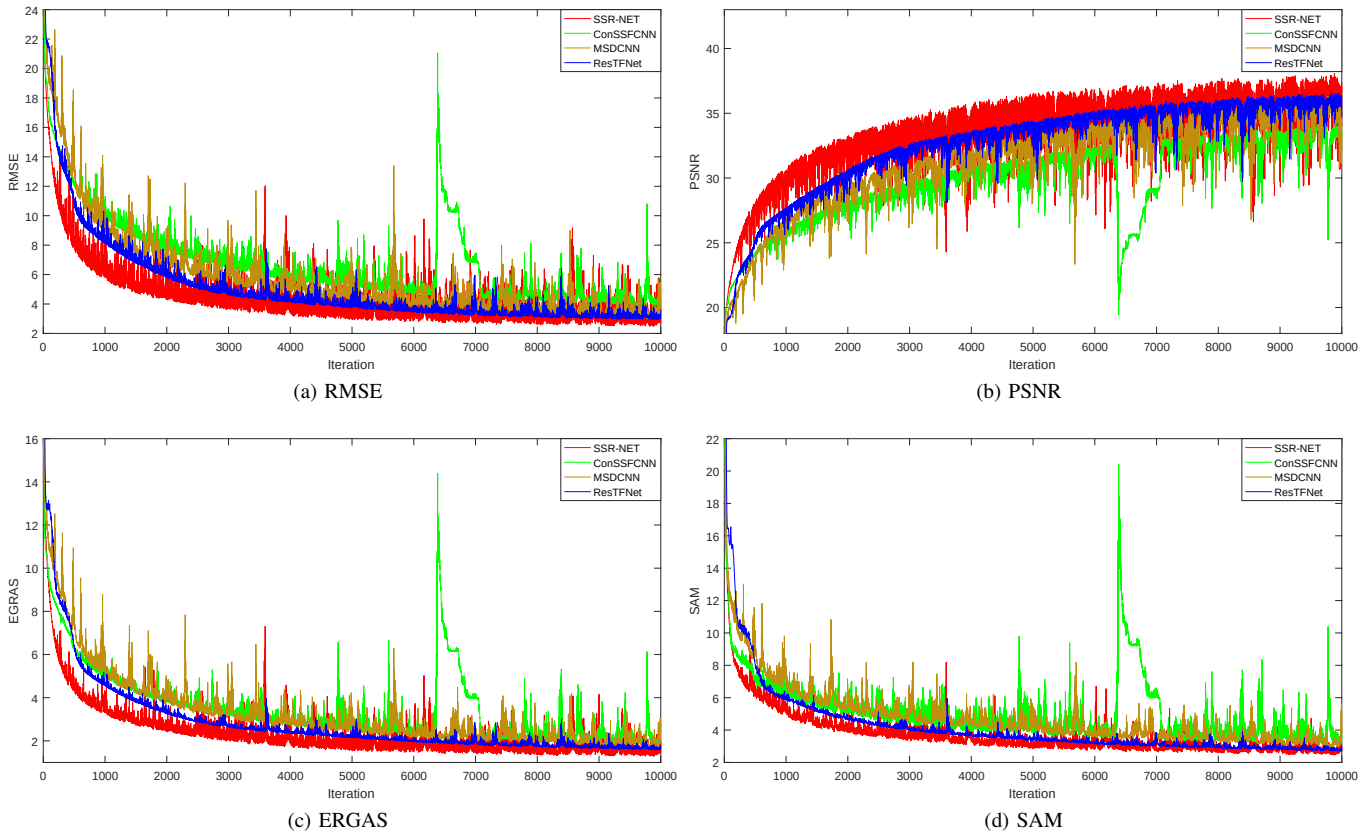
For Indian Pines (IP) dataset which is limited by the spatial resolution, the center 64 $\times$ 64 sub-region is cropped as the test image, and the rest region is used for training. More specifically, in each iteration, the training image with the same spatial resolution of 64 $\times$ 64 is randomly cropped from the training region. For all the other five datasets, the center 128 $\times$ 128 sub-region is cropped as the test image, and the rest region is used for training. Similarly, in each iteration, the training image with the same spatial resolution of 128 $\times$ 128 is randomly cropped from the training region. It is worth mentioning that the training and test region are non-overlapping, which is achieved by padding the test region with zeros in the datasets during the training stage. An example of the train and test regions of Urban dataset is illustrated in Fig. 4. The LR-HSI are down-sampled with the ratio of r set to 4 from the blurred HR-HSI, which is in advance processed by 5 $\times$ 5 Gaussian filter with the standard deviation set to 2. The HR-MSI is composed of the five images, which are located in HR-HSI at equal intervals.

Two traditional methods of CNMF [12] and LTTR [25], and five deep learning methods of TFNet [37], ResTFNet [37], SSFCNN [38], ConSSFCNN [38] and MSDCNN [39] are selected as the comparison approaches to evaluate the performance of the proposed SSR-NET. For the traditional methods, except data processing, all the parameters are set as the same as the original literatures. For all the deep learning models, the channel number of its input and output are adaptive to the used dataset. During the training stage, Adam is selected as the optimizer. All the models are trained for 10,000 iterations with the learning rate of 1e-4. It is worth mentioning that since the training images are randomly cropped from the training area, in this paper, iteration is used as the interval of metric measurement instead of epoch.

Besides, all the deep learning based experiments are implemented by Pytorch 1.3.0 on Python 3.7. The computing equipment contains the 64GB CPU memory and 1$\times$ GPU of GeForce GTX 1080Ti.

D. Ablation Study of SSR-NET

In order to explore the role of the components of SSR-NET in hyperspectral and multispectral image fusion, some ablation experiments are conducted on the dataset of Urban



(a) RMSE (b) PSNR

(c) ERGAS (d) SAM

Fig. 12: The metrics between the proposed SSR-NET and three comparison methods of MSDCNN, ConSSFCNN and ResTFNet on Urban dataset during the training stage.

TABLE II: The ablation experimental results of the proposed SSR-NET on Urban dataset. The arrow attached to the metrics points to the better performance. The best scores are marked in **bold**.

Model	Urban			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
Spat-CNN	18.1819	20.6904	9.8345	11.5203
Spec-CNN	7.7957	28.0460	3.8693	7.8680
SpatR-NET	15.5343	22.0573	8.5497	9.9719
SpecR-NET	7.0083	28.9709	3.3848	5.1665
SSR-NET	**2.3693**	**38.3909**	**1.2173**	**2.3124**

based on the five model variations of Spat-CNN, Spec-CNN, SpatR-NET, SpecR-NET and SSR-NET.

The results of ablation experiments are shown in Table II. Although SpatR-NET and SpecR-NET have the similar network architectures with Spat-CNN and Spec-CNN, the first two models perform quite better when compared to the last two models. The reason is probably that the inputs of SpatR-NET and SpecR-NET are both HMSI which contains more information than a single LR-HSI or HR-MSI, and more valuable spatial/spectral feature can be learned under the guidance of the spatial edge loss of $\mathcal{L}_{spat}$ and the spectrul edge loss of $\mathcal{L}_{spec}$. When comparing Spec-CNN and SpecR-NET with Spat-CNN and SpatR-NET, it could be found that the first two models reconstructing the spectral information have the better performance than the last two models reconstructing the spatial information. Such results indicate that spectral information is easier to be reconstructed than spatial information. In other words, the spatial information is much more complicated than the spectral information. After combining the advantages of SpatR-NET and SpecR-NET, SSR-NET achieves the best image fusion performance, which is much better than each individual of them.

E. Comparison with State-of-the-Art Methods

To verify the effectiveness of the proposed SSR-NET, comparison experiments of the datasets of Urban, PU, PC and Botswana are performed on seven state-of-the-art approaches, including CNMF [12], LTTR [25], TFNet [37], ResTFNet [37], SSFCNN [38], ConSSFCNN [38] and MSDCNN [39]. CNMF is a physical straight-forward matrix factorization based method, which aims to utilize the hyperspectral end-member matrix and the high-spatial-resolution abundance matrix to simulate the HR-HSI based on LR-HSI and HR-MSI. LTTR algorithm is a tensor-factorization based method, which designs an LTTR prior to learn the relationships among spectral, spatial, and nonlocal modes, and correspondingly proposes a low tensor train (TT) rank (LTTR)-based HSI super-resolution approach. TFNet, ResTFNet, SSFCNN, Con-

TABLE III: The comparison results of the proposed SSR-NET and seven state-of-the-art methods on Urban dataset. The best scores are marked in red, and the second scores are marked in green.

Method	Urban			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
CNMF [12]	4.0198	36.0468	1.5612	2.4971
LTTR [25]	24.6837	20.2826	11.7672	9.1336
MSDCNN [39]	3.1317	35.9676	1.7652	3.1199
TFNet [37]	3.0405	36.2243	1.7146	2.8524
ResTFNet [37]	2.8916	36.6604	1.5980	2.7355
SSFCNN [38]	8.5801	27.2133	4.2196	8.7402
ConSSFCNN [38]	3.8841	34.0972	1.8671	3.1609
Our SSR-NET	2.3693	38.3909	1.2173	2.3124

TABLE IV: The comparison results of the proposed SSR-NET and seven state-of-the-art methods on Pavia University dataset. The best scores are marked in red, and the second scores are marked in green.

Method	Pavia University			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
CNMF [12]	7.5816	30.5356	3.8939	2.2469
LTTR [25]	9.4202	28.6496	5.5504	5.1883
MSDCNN [39]	2.3207	40.5032	1.6358	2.6309
TFNet [37]	2.2427	40.8004	1.6053	2.4890
ResTFNet [37]	2.0578	41.5477	1.5068	2.3517
SSFCNN [38]	1.9410	42.0550	1.3957	2.2121
ConSSFCNN [38]	2.3916	40.2420	1.6340	2.4235
Our SSR-NET	1.6447	43.4938	1.2146	1.9329

SSFCNN and MSDCNN belong to deep learning methods, but they have different emphases. TFNet and ResTFNet are two-stream networks that encode the spatial and spectral feature independently and then decode the HR-HSI using the fusion of spatial and spectral feature. Compared to TFNet, there are extra skip-connection operations in ResTFNet. SSFCNN and ConSSFCNN use the direct concatenated image of LR-HSI and HR-MSI to predict the HR-HSI, where the latter concatenates the HR-MSI in each convolutional layer. Based on residual learning and multiscale feature extraction, MSDCNN presents a multiscale and multidepth CNN for remote sensing image fusion.

1) Results on Urban: Firstly, the comparison experiments based on the proposed SSR-NET and the aforementioned seven comparative methods are conducted on Urban dataset and the experimental results are provided in Table III. In order to show the fusion performance intuitively, Fig. 6 illustrates the fusion performance of all the used methods with the fusion R-G-B images, which are selected from the estimated HR-HSI, and the difference images of the fusion R-G-B images and the corresponding reference R-G-B images. Among the deep learning methods, the performance of SSFCNN on Urban is not good enough, which falls far behind ConSSFCNN. It indicates that the direct concatenation between HR-MSI and LR-HSI↑ is not suitable for Urban dataset.

TABLE V: The comparison results of the proposed SSR-NET and seven state-of-the-art methods on Pavia Center dataset. The best scores are marked in red, and the second scores are marked in green.

Method	Pavia Center			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
CNMF [12]	13.9777	25.2221	11.4360	4.3635
LTTR [25]	28.3373	19.0836	31.1229	17.0800
MSDCNN [39]	4,1687	35.7307	4.6840	5.4004
TFNet [37]	4.2473	35.5686	4.8295	4.8437
ResTFNet [37]	3.9258	36.2522	4.4449	4.6073
SSFCNN [38]	4.7720	34.5569	5.7509	5.7962
ConSSFCNN [38]	5.4998	33.3239	6.1594	6.1196
Our SSR-NET	3.4102	37.4752	3.8964	3.8618

TABLE VI: The comparison results of the proposed SSR-NET and seven state-of-the-art methods on Botswana dataset. The best scores are marked in red, and the second scores are marked in green.

Method	Botswana			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
CNMF [12]	26.3457	19.7166	9.4849	2.4866
LTTR [25]	9.1525	28.9000	14.2396	4.8682
MSDCNN [39]	0.5544	36.3497	3.1753	2.5996
TFNet [37]	0.4991	37.2622	2.7884	2.3161
ResTFNet [37]	0.4598	37.9749	2.6952	2.1537
SSFCNN [38]	1.4151	28.2108	11.4498	7.0424
ConSSFCNN [38]	2.5058	23.2475	15.2048	12.9631
Our SSR-NET	0.4106	38.9583	2.7994	2.0181

Fig. 12 shows all the four evaluation metrics of the proposed SSR-NET and comparitive methods on Urban dataset during the training process. In the comparative methods, TFNet and SSFCNN are not shown in the figure due to the limit of space, and their better versions of ResTFNet and ConSSFCNN are enough for comparison. It can be seen that although the proposed SSR-NET has a larger fluctuation with noise, there is an obvious advantage of the convergence speed in SSR-NET when comparing with other methods, especially in the early training iterations from 0 to 5,000. It mainly benefits from the proposed loss of $\mathcal{L}_{spec}$ and $\mathcal{L}_{spat}$. And it can be also observed that our SSR-NET has the superior upper limits of performance on the used metrics, which can be found in the iterations from 5,000 to 10,000.

2) Results on Pavia University (PU): Secondly, Table IV shows the experimental results on PU dataset, and Fig. 7 shows the fusion R-G-B images and the different images. In Table IV, it can be easily seen that for all four evaluation metrics, the proposed SSR-NET achieves the best performance among all the methods with the obvious advantage for PU dataset. SSFCNN also obtains good performance on Pavia University dataset. In contrast, ConSSFCNN performs worse than its simply verson of SSFCNN, which may be caused by the unstable skip-concatenation that is harmful for training. And ResTFNet

TABLE VII: The comparison results of the proposed SSR-NET and seven state-of-the-art methods on Indian Pines dataset. The best scores are marked in red, and the second scores are marked in green.

Method	Indian Pines			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
CNMF [12]	10.2174	27.9440	3.2092	2.8280
LTTR [25]	44.1083	15.2404	226.8429	19.0087
MSDCNN [39]	4.2063	34.4827	2.5249	2.9217
TFNet [37]	4.2682	34.3559	2.1629	2.9586
ResTFNet [37]	3.9815	34.9598	2.0790	2.8160
SSFCNN [38]	12.0730	25.3245	12.2866	9.6967
ConSSFCNN [38]	7.4048	29.5705	15.5868	5.6114
Our SSR-NET	3.8475	35.2573	8.5339	2.7886

TABLE VIII: The comparison results of the proposed SSR-NET and seven state-of-the-art methods on Washington DC Mall dataset. The best scores are marked in red, and the second scores are marked in green.

Method	Washington DC Mall			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
CNMF [12]	57.7621	12.8979	1356.2909	32.6896
LTTR [25]	18.6020	22.7396	301.2300	13.3933
MSDCNN [39]	2.7328	36.7345	0.4778	0.8993
TFNet [37]	1.9704	39.5756	0.3460	0.6756
ResTFNet [37]	1.7348	40.6816	0.3035	0.5836
SSFCNN [38]	16.4613	21.1373	3.1689	6.8978
ConSSFCNN [38]	14.0913	22.4876	2.6805	5.8734
Our SSR-NET	2.0591	39.1928	0.3603	0.7201

performs better when comparing to TFNet, which is probably that the skip-connection in ResTFNet can extract more stable feature based on step-wise residual strategy. According to Fig. 7, deep learning methods achieve superior fusion performance than traditional methods. By taking advantages of both spatial information in HR-MSI and spectral information in LR-HSI, the proposed SSR-NET can obtain the best fusion quality in a physical straight-forward manner.

3) Results on Pavia Center (PC): Table V lists the experimental results of our SSR-NET and the same comparative methods on PC dataset, and Fig. 8 illustrates the fusion performance with the fusion R-G-B images of our SSR-NET and the comparative methods. In Table V, it can be found that for all four evaluation metrics, the proposed SSR-NET obtains the superior performance than all the other comparative methods. Whether from the perspective of spatial reconstruction quality (ERGAS), spectral reconstruction quality (SAM), or the element-wise reconstruction quality (RMSE and PSNR), the proposed SSR-NET has an obvious advantage. Similar to the results of PU dataset, The deep learning methods are much better than traditional methods. SSFCCN also performs worse than ConSSFCNN and ResTFNet still has better fusion quality than TFNet.

4) Results on Botswana: The experimental results of Botswana dataset on the proposed SSR-NET and the comparative methods are shown in Table VI, and the corresponding fusion R-G-B images are illustrated in Fig. 9. Among all the evaluation metrics, our SSR-NET obtains three best scores of RMSE, PSNR and SAM, and one competitive score of ERGAS. The first and second places of ERGAS are acquired by ResTFNet and TFNet. In Botswana dataset, the spatial resolution of a pixel is up to 30 m and thus its spatial information is much more complex than other datasets with the higher requirement for feature extraction. TFNet and ResTFNet, which are much deeper than our three-layer SSR-NET, have the advantage in extracting non-linear deep feature that is benefical for spatial reconstruction. It is probably that the proposed SSR-NET will perform better than TFNet and ResTFNet if SSR-NET has the same model size as TFNet and ResTFNet.

TABLE IX: The metric scores measured between the reference HR-HSI and the outputs of different stages in SSR-NET. The best scores are marked in **bold**.

Output	Urban			
	RMSE↓	PSNR↑	ERGAS↓	SAM↓
HMSI	14.5836	22.6058	7.9786	7.0940
Spectral HR-HSI	2.5869	37.6276	1.3295	2.6184
Spatial HR-HSI	**2.3693**	**38.3909**	**1.2173**	**2.3124**

5) Results on Indian Pines (IP): Table VII lists the comparison results of IP dataset between the proposed SSR-NET and the comparative methods, and Fig. 10 shows the corresponding fusion R-G-B images. It can be seen that our SSR-NET performs best among all the approaches in terms of RMSE, PSNR and SAM. As for ERGAS, the proposed SSR-NET achieves comparitive performance when comparing with other state-of-the-art methods. There are two potential factors leading to such the result. Firstly, it is probably that the metric of ERGAS concentrates on global information. Instead, the proposed Spatial Edge loss and Spectral Edge loss stress more on the local features, ignoring the global features. Secondly, limited by the model size, SSR-NET has a disadvantage in a large receptive field.

6) Results on Wanshing DC Mall (WDCM): The results of the proposed SSR-NET and the seven comparative methods on WDCM dataset are shown in Table VIII, and the fusion results of the used approaches on this dataset are illustrated in Fig. 11. It can be seen that similar to TFNet, the experimental results of our SSR-NET are in second echelon, which just slightly fall behind ResTFNet. It is probably that compared with TFNet and ResTFNet, the model size of the propsed SSR-NET is rather smaller, which limits the performance of our model on spatial feature extraction of a large receptive field. It is also our future work in hyperspectral and multispectral image fusion.

Overall, the proposed SSR-NET has the best fusion performance on the five datasets including Urban, PU, PC, Botswana and IP, and achieves the comparable results in WDCM dataset. Traditional methods of CNMF and LTTR perform worse than deep learning methods. Among deep learning methods, when compared to SSFCNN and ConSSFCNN, our SSR-NET has more stable performance. When compared to TFNet,

ResTFNet and MSDCNN, SSR-NET has not only better performance but also much smaller model size.

F. Analysis of the Interpretability of SSR-NET

Although deep learning methods obtain excellent performance, as black-box models, they are usually lack of interpretability. Different from the existing deep learning methods used in LR-HSI and HR-MSI fusion, the proposed SSR-NET is interpretable to some extent. Due to the Cross-Mode Message Inserting, all the outputs of different stages in SSR-NET, which have the same size as the reference HR-HSI, can serve as the estimated HR-HSI with defferent qualities. To explore the process of the fusion HR-HSI in the proposed SSR-NET, the outputs of different stages of Urban dataset are quantitatively measured by the above four evaluation metrics. The experimental results are shown in Table IX.

In the table, HMSI denotes the concatenation of LR-HSI and HR-MSI according to Eqn (4). Spectral HR-HSI and Spatial HR-HSI represent the output of Spectral Reconstruction Network and Spatial Reconstruction Network respectively. It could be found that the fusion qualities of HMSI, Spectral HR-HSI and Spatial HR-HSI are increasing, which demonstrates that the proposed SSR-NET is a physical straight-forward model.

TABLE X: Model size, FLOPs and test time of the proposed SSR-NET and the other deep learning models. The best data are marked in **bold**.

Model	Urban		
	Params (M)	FLOPs (G)	Test Time (Ms)
MSDCNN [39]	1.82	59.76	438
TFNet [37]	2.50	18.3	215
ResTFNet [37]	2.38	**17.04**	206
SSFCNN [38]	1.14	37.24	232
ConSSFCNN [38]	1.16	38.16	268
Our SSR-NET	**0.69**	23.26	**151**

G. Analysis of Model Complexity

For deep learning methods, model paramters and test time are also the important indicators for performance evaluation. In this sub-section, the number of model parameters, its floating point operations (FLOPs) and its test time of the proposed SSR-NET and the other five deep learning methods are compared quantitatively, and the detailed data are provided in Table X. The experiments are still performed on the test area of Urban dataset with using only CPU.

According to the results, it is easily seen that the proposed SSR-NET has an obvious advantage in both model parameters and test time in comparison with the other deep learning models. Compared to the TFNet, the parameters of our SSR-NET is only 28% of TFNet but the test speed is about 50% faster than it. As for FLOPs, the performance of our SSR-NET just slightly falls behind TFNet and ResTFNet. It is probably that all layers of our SSR-NET are operated at the pixel level. Taking the fusion quality into consideration again, the proposed SSR-NET has the overall advantage in the field of LR-HSI and HR-MSI fusion.

V. Conclusion

In this paper, an interpretable CNN-based framework of SSR-NET is proposed for hyperspectral and multispectral image fusion. The proposed SSR-NET contains three components of Cross-Mode Message Inserting, Spatial Reconstruction Network (SpatRN) and Spectral Reconstruction Network (SpecRN). SpatRN and SpecRN are respectively optimized by two kinds of losses of Spatial Edge Loss ($\mathcal{L}_{spat}$) and Spectral Edge Loss ($\mathcal{L}_{spec}$), which are specially designed for the spatial and spectral restoration. Comparative experiments of the proposed SSR-NET and some state-of-the-art methods are conducted on six widely used hyperspectral datasets, including Urban, Pavia University, Pavia Center and Botswana. The superior experimental results of SSR-NET demonstrate the effectiveness of the proposed method. Besides the scores of evaluation metrics, our SSR-NET also has an obvious advantage in model size and test time in the type of deep learning methods.

References

[1] F. Luo, L. Zhang, B. Du, and L. Zhang, "Dimensionality reduction with enhanced hybrid-graph discriminant learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[2] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[3] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 911–923, 2018.

[4] H. Yan, Y. Zhang, W. Wei, L. Zhang, and Y. Li, "Salient object detection in hyperspectral imagery using spectral gradient contrast," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 1560–1563.

[5] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5910–5922, 2018.

[6] W. Sun and Q. Du, "Graph-regularized fast and robust principal component analysis for hyperspectral band selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3185–3195, 2018.

[7] Y. Yuan, X. Zheng, and X. Lu, "Discovering diverse subset for unsupervised hyperspectral band selection," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 51–64, 2016.

[8] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, pp. 1279–1289, 2016.

[9] D. Marinelli, F. Bovolo, and L. Bruzzone, "A novel change detection method for multitemporal hyperspectral images based on binary hyperspectral change vectors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4913–4928, 2019.

[10] Q. Wang, Z. Yuan, Q. Du, and X. Li, "Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2018.

[11] J. Zhou, C. Kwan, B. Ayhan, and M. T. Eismann, "A novel cluster kernel rx algorithm for anomaly and change detection using hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, pp. 6497–6504, 2016.

[12] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2011.

[13] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3586–3594.

[14] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, and X. Li, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2337–2352, 2016.

[15] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3631–3640.
[16] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5344–5353.
[17] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE transactions on neural networks and learning systems*, no. 99, pp. 1–11, 2018.
[18] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.
[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[21] C.-H. Lin, F. Ma, C.-Y. Chi, and C.-H. Hsieh, "A convex optimization-based coupled nonnegative matrix factorization algorithm for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1652–1667, 2017.
[22] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a sylvester equation," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4109–4121, 2015.
[23] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.
[24] K. Zhang, M. Wang, S. Yang, and L. Jiao, "Spatial–spectral-graph-regularized low-rank tensor decomposition for multispectral and hyperspectral image fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 4, pp. 1030–1040, 2018.
[25] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2672–2683, 2019.
[26] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[28] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2020.
[29] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 10 039–10 042.
[30] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected topics in applied earth observations and remote sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.
[31] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.
[32] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3516–3530, 2017.
[33] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, 2017.
[34] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by ms/hs fusion net," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1585–1594.
[35] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by cnn denoiser," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
[36] S. Xu, O. Amira, J. Liu, C.-X. Zhang, J. Zhang, and G. Li, "Ham-mfn: Hyperspectral and multispectral image multiscale fusion network with rap loss," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
[37] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Information Fusion*, vol. 55, pp. 1–15, 2020.
[38] X.-H. Han, B. Shi, and Y. Zheng, "Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2506–2510.
[39] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978–989, 2018.
[40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
[41] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" 2000.
[42] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm," 1992.

Xueting Zhang received the B.E. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018. She is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. Her research mainly focuses remote sensing image proccessing.

Wei Huang received the B.E. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and remote sensing.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Xuelong Li (M'02-SM'07-F'12) is currently a Professor with the School of Computer Science, with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China.