SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2022, Vol. 65 152102:1–152102:14
https://doi.org/10.1007/s11432-020-3102-9

Learning hyperspectral images from RGB images via a coarse-to-fine CNN

Shaohui MEI[1*], Yunhao GENG[1], Junhui HOU[2] & Qian DU[3]

[1]*School of Electronics and Information, Northwestern Polytechnical University, Xi'an* 710072*, China;*
[2]*Department of Computer Science, City University of Hong Kong, Hong Kong* 999077*, China;*
[3]*Department of Electrical and Computer Engineering, Mississippi State University, Starkville* MS 39762*, USA*

Received 1 April 2020/Revised 14 July 2020/Accepted 29 September 2020/Published online 6 September 2021

Abstract Hyperspectral remote sensing is well-known for its extraordinary spectral distinguishability to discriminate different materials. However, the cost of hyperspectral image (HSI) acquisition is much higher compared to traditional RGB imaging. In addition, spatial and temporal resolutions are sacrificed to obtain very high spectral resolution owing to the limitations of sensor technologies. Therefore, in this paper, HSIs are reconstructed using easily acquired RGB images and a convolutional neural network (CNN). As a result, high spatial and temporal resolution RGB images can be inherited to HSIs. Specifically, a two-stage CNN, referred to as the spectral super-resolution network (SSR-Net), is designed to learn the transformation model between RGB images and HSIs from training data, including a band prediction network (BP-Net) to estimate hyperspectral bands from RGB images and a refinement network (RF-Net) to further reduce spectral distortion in the band prediction step. As a result, the learned joint features in the proposed SSR-Net can directly predict HSIs from their corresponding scenes in RGB images without prior knowledge. Experimental results obtained on several benchmark datasets demonstrate that the proposed SSR-Net outperforms several state-of-the-art methods by ensuring higher quality in HSI reconstruction, and significantly improves the performance of traditional RGB images in classification.

Keywords hyperspectral, reconstruction, convolutional neural network, deep learning

Citation Mei S H, Geng Y H, Hou J H, et al. Learning hyperspectral images from RGB images via a coarse-to-fine CNN. Sci China Inf Sci, 2022, 65(5): 152102, https://doi.org/10.1007/s11432-020-3102-9

1 Introduction

Hyperspectral imaging technology, which collects electromagnetic spectrum information in hundreds of narrow and contiguous bands, has been widely used in many applications, e.g., land-use/land-cover mapping, mineral exploration, and water pollution detection [1–3]. To collect rich information about an imaged scene, hyperspectral imaging systems must capture a large amount of spatial, spectral, and temporal information. However, compared to high spectral resolution, the spatial and temporal resolutions of hyperspectral images (HSI) are not as high as that of traditional RGB and multispectral images because they are balanced with spectral resolution owing to the limitations of optical systems and sensor technologies.

RGB imaging can easily achieve extremely high spatial and temporal resolutions at very low acquisition costs. As a result, RGB images are used in an extremely wide range of applications. The abundant structure and texture information contained in RGB images is very relative to identifying objects and understanding scenes. In addition, the high temporal resolution of RGB videos can enable tracking of moving objects and understanding human behavior. However, compared to HSIs, the lack of spectral information in RGB images degrades the ability to discriminate pixels of different materials.

In many applications, images (or videos) with high spatial, spectral, and temporal resolutions are required to identify and track objects according to their structure, texture, and spectra [4–8]. However, owing to the trade-offs among spatial, spectral, and temporal resolutions in an imaging system, it is

* Corresponding author (email: meish@nwpu.edu.cn)

very difficult to directly acquire images with high resolution in these domains. Therefore, many post-acquisition prepossessing algorithms have been proposed. For example, image fusion algorithms fuse spectral information in low spatial-resolution HSIs and spatial information in high spatial-resolution multispectral images, mosaic RGB images, and panchromatic images [9–14], where images with both high spatial and spectral resolution can be obtained. The spatial super resolution (SR) algorithm can generate a super-resolution image using a low spatial-resolution HSI [15–18]. However, the imaging cost is high owing to the input of HSIs. In addition, the temporal resolution in these strategies cannot be improved further owing to the temporal resolution limitations of HSIs. Therefore, to reduce image acquisition costs, spectral SR of RGB images has been proposed to obtain HSIs with high spatial-resolution. As a result, the temporal resolution of such HSIs can be enhanced because RGB images can be acquired easily at very high temporal resolution. In addition, spatial and spectral joint SR has also been attempted over existing satellite multispectral images to obtain high spatial-resolution HSIs [19].

Recently, some studies have explored direct hyperspectral reconstruction based on RGB images. For example, Nguyen et al. [20] employed a radial basis function (RBF) neural network to estimate scene reflections and global illumination. However, the color matching function was assumed to be known, and its reconstruction performance was highly dependent on the white balance step. Arad et al. [21] studied a sparse dictionary between an HSI and its corresponding RGB image, and they reconstructed an HSI based on the sparse reconstruction. Yi et al. [22] improved Arad's method using a spectral improvement and spatial preservation strategy. Jia et al. [23] assumed that hyperspectral pixels can be embedded in low-dimensional manifolds, where the low-dimensional manifolds are reconstructed based on an RBF neural network. As a result, HSIs are reconstructed from the learned low-dimensional manifolds, rather than being learned from the original RGB images. In 2018, CVPR organized a challenge [24] on spectral reconstruction from RGB images, where an advanced CNN-based hyperspectral reconstruction method achieved the best performance. Since then, many CNN-based methods have been proposed. For example, Can and Timofte [25] proposed a moderately deep CNN model to reconstruct spectral images. Han et al. [26] employed class-based back propagation neural networks to learn nonlinear spectral mappings between RGB and high-spatial-resolution HSI pairs. In addition, 2D and 3D CNNs have been applied to spectral reconstruction from RGB images, in which 3D-CNN based architecture achieved better performance because it can exploit the inter-channel co-relation to refine the extraction of spectral data [27].

HSI datasets acquired by well-known airborne sensors (e.g., AVIRIS) and other natural scenes on the ground (e.g., The Interdisciplinary Computational Vision Laboratory (ICVL) datasets [21]) have provided abundant raw materials for data-based learning. Based on big data resources, transfer learning enables the reconstruction of spectral information in scenes from RGB images and learning the complex relationships between the two data sources. Deep learning technologies, which have made remarkable progress in feature learning tasks, can effectively exploit the advantages of big data to learn the spatial context features of complex scenes from HSIs [28–31]. Therefore, in this paper, we propose a spectral super-resolution network (SSR-Net), which applies deep learning to realize spectral SR of RGB images. The proposed SSR-Net contains a band prediction network (BP-Net) to increase spectral resolution of RGB images and a refinement network (RF-Net) to fine-tune spectral distortion in BP-Net. In the proposed BP-Net, only the convolutional layer is used, where the common characteristics of all bands in HSIs are learned jointly in the first few convolutional layers, and different band images are reconstructed with different final convolutional layers and different convolution kernels. In all convolutional layers (other than the first layer), a small convolution kernel, i.e., $3\times3\times3$, is used to increase learning by considering a small number of parameters. In the proposed RF-Net, dilate convolution is employed to learn the correlation between non-continuous bands such that more spectral features can be extracted. In addition, a residual branch is added to increase the training speed and improve reconstruction performance. Finally, experiments conducted on three well-known public datasets (the ICVL [21], CAVE (The CAVE Laboratory Multispectral Image Database) [32], and KAIST (Visual Computing Laboratory of the KAIST University) [33] datasets) were conducted to compare the proposed SSR-Net to state-of-the-art spectral SR algorithms. In addition, experiments on HSIs were performed to demonstrate the advantages of spectral SR in classification tasks.

In summary, our primary contributions of this paper are summarized as follows.

(1) We proposed a two-stage neural network (SSR-Net) to learn the mapping between HSIs and RGB images, where HSIs can be reconstructed directly from RGB inputs without prior knowledge in a coarse-to-fine manner. The proposed SSR-Net is designed according to the unique characteristics of HSIs, including high spectral similarity among band images and abundant spatial context. Since band images

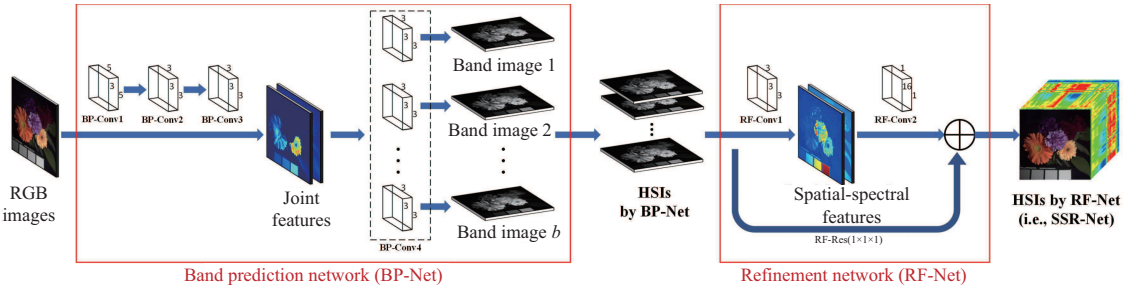

Figure 1 (Color online) Framework of proposed SSR-Net for HSI reconstruction. The proposed SSR-Net estimates HSIs from RGB images in a coarse-to-fine manner. First, the BP-Net estimates HSIs in a coarse manner, where joint spatial-spectral features are learned, and then all band images are predicted in a band-by-band manner. The RF-Net further fine-tunes HSIs by exploiting spatial context and spectral correlation in the coarsely estimated HSIs.

in an HSI have high spectral similarity, joint spatial-spectral features are first learned in BP-Net to estimate HSIs in a band-by-band manner for coarse estimation. To facilitate fine estimation, the RF-Net is constructed to fine-tune the reconstructed results by further exploiting the spatial context and spectral correlation in the coarsely estimated HSIs.

(2) In addition to validation using natural scenes, i.e., the ICVL, CAVE, and KAIST datasets, the proposed SSR-Net is also evaluated using satellite HSIs, and the superiority of spectral SR is validated in the context of classification applications. The experimental results demonstrate the spectral SR can improve the classification performance of remote sensing images by reconstructing HSIs from the original RGB inputs.

The remainder of this paper is organized as follows. Section 2 presents the proposed SSR-Net for hyperspectral reconstruction, including BP-Net for band image prediction and RF-Net for spectral fine tuning. In Section 3, we discuss experiments conducted on several benchmark datasets to demonstrate the effectiveness of the proposed SSR-Net for hyperspectral reconstruction. Finally, the conclusion is presented in Section 4.

2 Proposed SSR-Net

Notation. In this paper, scalars are denoted by italic lowercase letters, vectors by bold lowercase ones, 2-D matrices by bold uppercase ones, three-dimensional (3-D) and higher dimensional matrices by calligraphy ones, and functions/operators by script ones. Let $\mathcal{R} \in \mathbb{R}^{m\times n\times 3}$ denote an RGB image with $m \times n$ pixels. Correspondingly, let $\mathcal{H} \in \mathbb{R}^{m\times n\times b}$ represent the b-band HSI reconstructed from $\mathcal{R}$, where $\mathbf{I}^{(i)} \in \mathbb{R}^{m\times n}$ $(1 \leqslant i \leqslant b)$ represents the i-th spectral band of the reconstructed HSI.

In hyperspectral remote sensing, tens or hundreds of images are acquired simultaneously, owing to which the acquisition cost is high and spatial resolution is compromised for high spectral resolution. Moreover, it is very difficult to acquire hyperspectral videos. On the contrary, RGB images can be acquired at a very low cost with very high spatial and temporal resolution. Therefore, in this paper, RGB images are used to reconstruct HSIs using deep learning techniques, by which the cost of HSI acquisition will be greatly reduced. The reconstructed HSIs own very high spatial resolution as the input RGB images. Moreover, the temporal resolution of HSIs can also be enhanced when RGB videos (or image sequences) are used as input. The entire deep learning framework of our proposed SSR-Net for HSI reconstruction is shown in Figure 1, which shows the proposed SSR-Net contains two parts: a BP-Net to separately reconstruct all the bands of HSIs and an RF-Net to reduce spectral distortion in previous prediction.

2.1 Proposed BP-Net

The BP-Net is first used to reconstruct all the hyperspectral band images directly from the input RGB image without any auxiliary information. As shown in Figure 1, the proposed BP-Net utilizes fully convolutional layers to conduct such a prediction task. Different from a traditional image prediction network that outputs just a single band image, the proposed BP-Net predicts tens or hundreds of hyperspectral band images simultaneously. Instead of directly constructing tens or hundreds of isolated networks, the proposed BP-Net first extracts joint features from RGB images for all the band predictions. Then multiple convolution kernels layer is designed to predict different band images from these extracted joint

Table 1 The structure parameters of the proposed SSR-Net

		Input	Kernel	Padding	Dilation	Output
BP-Net	BP-Conv1	1, 3, 64, 64	32, 1, 3, 5, 5	1, 2, 2	–	32, 3, 64, 64
	BP-Conv2	32, 3, 64, 64	32, 32, 3, 3, 3	1, 1, 1	–	32, 3, 64, 64
	BP-Conv3	32, 3, 64, 64	32, 32, 3, 3, 3	1, 1, 1	–	32, 3, 64, 64
	BP-Conv4 ($\times b$)	32, 3, 64, 64	32, 3, 3, 3 ($\times b$)	0, 1, 1	–	1, 1, 64, 64 ($\times b$)
RF-Net	RF-Conv1	1, b, 64, 64	16, 1, 3, 3, 3	3, 1, 1	3, 1, 1	16, b, 64, 64
	RF-Conv2	16, b, 64, 64	1, 16, 1, 1, 1	–	–	1, b, 64, 64
	RF-Res	1, b, 64, 64	1, 1, 1, 1, 1	–	–	1, b, 64, 64

features. As a result, the following two phases are contained in the proposed BP-Net.

2.1.1 *Joint feature learning*

This phase uses three successive 3-dimensional (3D) convolutional layers to learn effective features from the RGB input, which are jointly used to predict band images. Such 3D convolution simultaneously explores the spatial context and spectral information within the three RGB channels. Each convolutional layer nonlinearly transforms its input $\mathcal{F}^{(i)}$ ($i = 0, 1, 2$ and $\mathcal{F}^{(0)} = \mathcal{R}$) into multi-channel feature maps denoted as $\mathcal{F}^{(i+1)} \in \mathbb{R}^{32\times3\times64\times64}$:

$$\mathcal{F}^{(i)} = f\left(\mathcal{W}_J^{(i)} \otimes \mathcal{F}^{(i-1)} + \delta_J^{(i)}\right), \quad i = 1, 2, 3, \tag{1}$$

where $\mathcal{W}_J^{(i)}$ and $\delta_J^{(i)}$ respectively represent the convolutional kernel and bias in this three convolutional layer, $f(\cdot)$ represents the activation function in these convolutional layers, and '$\otimes$' represents the convolution operation. In the SSR-Net, the rectified linear unit (Relu) function is adopted, which is conducted in element-wise mode as

$$f(x) = \begin{cases} x, & \text{if } x > 0, \\ \alpha x, & \text{if } x \leqslant 0, \end{cases} \tag{2}$$

with α being a trainable parameter involved in this activation function. Note that padding is used in all convolutional layers to expand the edges of images/features to ensure that the output of each layer of the feature maps has an identical size to its input.

2.1.2 *Band image prediction*

This phase separately estimates all the band images of an HSI from the learned joint features $\mathcal{F}^{(3)}$ using convolutional layer. One convolutional layer is constructed to estimate just one band image and totally b convolutional layers are required to estimate all the band images. The j-th band image of HSI is estimated as

$$\mathbf{I}^{(j)} = \mathcal{W}_P^{(j)} \otimes \mathcal{F}^{(3)} + \delta_P^{(j)}, \quad j = 1, 2, \ldots, b, \tag{3}$$

where $\mathcal{W}_P^{(j)}$ and $\delta_P^{(j)}$ respectively represent the convolutional kernel and bias to estimate the j-th band image. Note that no activation function is applied in these convolutional layers.

In a CNN, the size of the mapped region of pixels on the feature map output plays an important role in feature learning. As shown in Figure 2(c), the size of the receptive field of the first convolution layer equals the size of the filter, while that of deeper convolutional layer is related to both the stride size and the size of convolution kernel of all previous layers. Thus, a small convolution kernel can also result in a large receptive field under multi-layer superposition [34]. Moreover, a small convolution kernel offers the advantage of fast learning speed since fewer parameters are involved. Therefore, in the proposed BP-Net, small convolution kernels of size $3 \times 3 \times 3$ are used for learning in all the convolutional layers except the first layer that is to reduce the number of parameters and accelerate the learning speed. The detailed information of the proposed BP-Net is listed in Table 1, which shows that the size of input to the BP-Net, denoted as $\mathcal{R}$, is fixed as $1 \times 3 \times 64 \times 64$, meaning an RGB image of size 64×64 is fed to the BP-Net. According to the information of the four convolutional layers in the proposed BP-Net listed in Table 1, the size of receptive field for these four layers is shown in Table 2. It can be observed that, though small convolution kernels are used in the BP-Net, areas as large as 11×11 from the RGB image can be sensed to estimate just one element for the HSI.

Table 2 The receptive field of each convolutional layer in the BP-Net

	Conv1	Conv2	Conv3	Conv4
Receptive field	5×5	7×7	9×9	11×11
Kernel	$32, 3, 5, 5$	$32, 3, 3, 3$	$32, 3, 3, 3$	$31, 3, 3, 3$

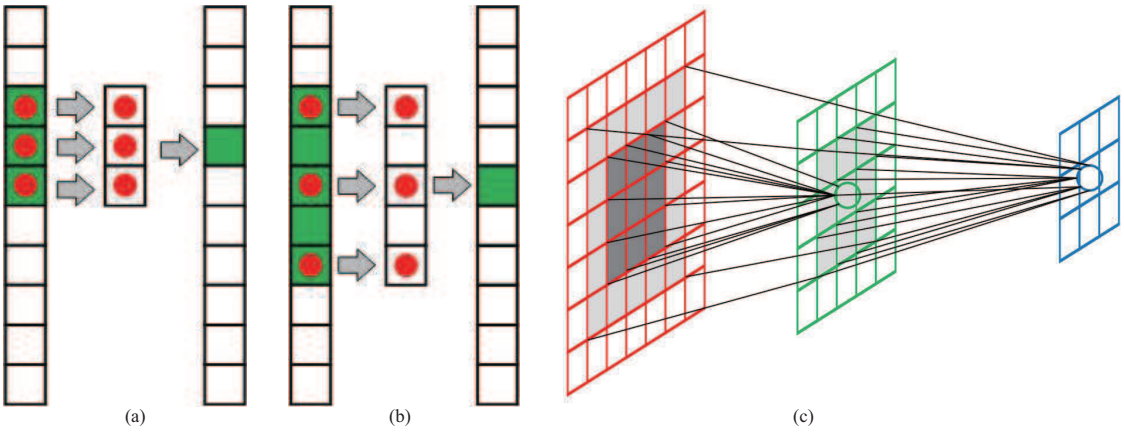

Figure 2 (Color online) (a) Illustration of a traditional convolutional layer. (b) Illustration of a dilate convolutional layer. In these figures, red indicates the region of the input layer that is dotted with the convolution kernel, while green indicates the range of receptive fields corresponding to the convolution kernel. (c) Illustration of the receptive field. In CNNs, the receptive field of a neuron is determined by the filter in all the preceding layers.

2.2 Proposed RF-Net

In the proposed BP-Net, spectral distortion easily occurred since all the bands of an HSI are estimated independently. Therefore, another RF-Net is constructed to alleviate the spectral distortion. Context information has played an important role in image reconstruction, image prediction and other tasks. In this paper, spectral context information is considered in the proposed RF-Net.

In order to alleviate spectral distortion, more contextual spatial-spectral information can be explored by increasing the depth of the network or the size of the convolution kernel. However, such an operation also results in difficulty in network learning since much more parameters are involved. Dilate convolution, which is also known as atrous convolution in DeepLab [35], allows for an exponential increase in the field of view without the decrease of spatial dimensions or increase of the number of parameters [36]. As shown in Figures 2(a) and (b), compared with traditional convolution, the dilate convolution learns features from non-continuous bands, such that a larger size of spectral context can be explored under a similar scale of network to traditional convolution. For example, if traditional convolution is applied on an input $\boldsymbol{x} \in \mathbb{R}^{p\times 1}$ with a kernel $\boldsymbol{w} \in \mathbb{R}^{q\times 1}$ (without loss of generality, $q \ll p$), the output $\boldsymbol{y} \in \mathbb{R}^{p\times 1}$ (padding can be used to keep the size of output equal to that of input) is

$$\boldsymbol{y}_k = \sum_{l=0}^{q-1} \boldsymbol{w}_l \boldsymbol{x}_{k+l}, \quad k = 1, 2, \ldots, p. \tag{4}$$

However, when the dilate convolution is applied, the output $\boldsymbol{y}' \in \mathbb{R}^{(p)\times 1}$ becomes

$$\boldsymbol{y}'_k = \sum_{l=0}^{q-1} \boldsymbol{w}_l \boldsymbol{x}_{k+l*\Delta}, \quad k = 1, 2, \ldots, p, \tag{5}$$

where Δ is a step-size in dilate convolution and set to 2 in this paper. High dimensional dilate convolution can be obtained by applying this one-dimensional calculation to all dimensions. Therefore, dilate convolution is adopted in the proposed RF-Net to alleviate spectral distortion by effectively exploring the spectral context of non-continuous bands.

As shown in Figure 1, the proposed RF-Net consists of two dilate convolutional layers to explore both spatial and spectral context. The first convolutional layer learns effective spatial-spectral features from the HSI estimated by the BP-Net, while the second convolutional layer estimates the HSI using

these features. The output of BP-Net, which is constructed by assembling these estimated band images $\mathcal{H}_B = \{\mathbf{I}^{(j)}\}$, is fed to the RF-Net. Therefore, the feature learned in the first convolutional layer is

$$\boldsymbol{F}_R = f\left(\mathcal{W}_{R1}\bar{\otimes}\mathcal{H}_B + \delta_{R1}\right), \tag{6}$$

where $\mathcal{W}_{R1}$ and δ_{R1} respectively represent the weight and bias in this convolution, $\boldsymbol{F}_R$ represents the features learned in RF-Net, and $\bar{\otimes}$ denotes dilate convolution. The Relu activation function is also used in this feature learning convolutional layer. Consequently, an updated HSI can be estimated from these features learned in $\mathcal{H}_B$:

$$\boldsymbol{H}_R = \mathcal{W}_{R2}\bar{\otimes}\mathcal{F}_R + \delta_{R2}, \tag{7}$$

where $\mathcal{W}_{R2}$ and δ_{R2} respectively represent the weight and bias in the second convolutional layer, $\boldsymbol{H}_R$ represents the estimated HSI after alleviating spectral distortion by exploring spatial-spectral context.

In order to accelerate learning processing, a residual branch is also adopted in the RF-Net, where a $1 \times 1 \times 1$ convolutional kernel is used for linear scaling to fine-tune the range of the output image. As a result, the final HSI estimated from the RGB image is

$$\hat{\boldsymbol{H}} = \boldsymbol{H}_R + \mathcal{W}_B \otimes \mathcal{H}_B, \tag{8}$$

where $\mathcal{W}_B$ represents the $1 \times 1 \times 1$ convolutional kernel in residual branch and $\hat{\boldsymbol{H}}$ represents the HSI estimated from the RF-Net, i.e., the HSI estimated from the proposed SSR-Net. The parameters of the proposed RF-Net are also summarized in Table 1.

2.3 Training procedure

The RGB input and its corresponding ground-truth HSI are used to train the network, and the L_1 loss function [37] is adopted in the proposed SSR-Net. Assumed $\boldsymbol{H}$ represents ground-truth HSI of the HSI estimated by the SSR-Net $\hat{\boldsymbol{H}}$. The L_1 loss function between the prediction and the ground-truth in the SSR-Net is defined as follows:

$$\text{loss}(\boldsymbol{H}, \hat{\boldsymbol{H}}) = \frac{1}{m \times n \times c} \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{k=1}^{c} |(h_{i,j,k} - \hat{h}_{i,j,k})|, \tag{9}$$

where $h_{i,j,k}$ and $\hat{h}_{i,j,k}$ represent the input at the position (i, j, k) of $\boldsymbol{H}$ and $\hat{\boldsymbol{H}}$, respectively, and '$|\cdot|$' represents absolute value.

The proposed SSR-Net estimates hyperspectral images (HSIs) from RGB images in a coarse-to-fine manner: BP-Net first estimates HSIs in a coarse manner, in which joint spatial-spectral features are first learned and then all the band images are predicted in a band by band manner; RF-Net further fine-tunes coarsely estimated HSIs by exploiting their spatial context and spectral correlation. Therefore, the training of the proposed SSR-Net can be divided into two steps: training BP-Net for coarse estimation and training RF-Net for fine estimation, When the proposed SSR-Net is well-trained, it can be directly used to reconstruct HSIs from an RGB inputs. Note that in order to train the proposed SSR-Net more effectively, the orthogonal method proposed in [38] is used for weight initialization. In this initialization algorithm, the weight of network is filled with a semi-orthogonal matrix to accelerate the training speed.

3 Experiments

In this section, extensive experiments are conducted to evaluate the performance of our proposed SSR-Net for the reconstruction of HSIs from RGB images.

3.1 Experiments on ground natural images

3.1.1 *Datasets*

Three benchmark hyperspectral datasets, namely ICVL dataset [21], CAVE dataset [32] and KAIST dataset [33], are adopted. The ICVL dataset is acquired using a Specim PS Kappa DX4 hyperspectral camera and a rotary stage for spatial scanning [21]. Such dataset contains 201 images from a variety of urban (residential/commercial), suburban, rural, indoor and plant-life scenes. Most images are of size

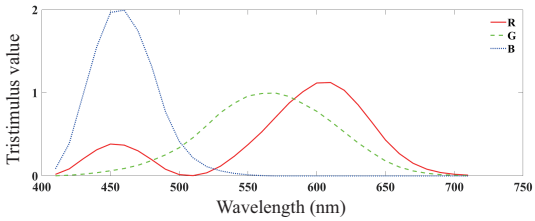

Figure 3 (Color online) CIE_1964 color match function.

Table 3 Reconstruction performance of the proposed SSR-Net with different training strategies over the ICVL dataset

	RMSE	PSNR (dB)	SSIM	SAM
Overall training	0.0072	43.88	0.9913	0.0387
Coarse-to-fine training	0.0063	45.12	0.9929	0.0337

1392×1300 and 519 spectral bands (from 400 nm to 1000 nm at roughly 1.25 nm increments). Similar to that in [21], the spectral range used from each image was limited roughly to the visual spectrum and computationally reduced via proper binning of the original narrow bands to 31 bands of roughly 10 nm in the range 400–700 nm. The CAVE dataset, which is available at the web[1)], is acquired using a tunable filter (VariSpec Liquid Crystal Tunable Filter) and a cooled CCD camera (Apogee Alta U260, 512×512 pixels) [32]. The images in this dataset contain 31 bands ranging from 400 nm to 700 nm with 10 nm intervals. A variety of objects and materials are included in this dataset, such as textiles, skin, hair, real and fake fruits and vegetables, candy, drinks, and paints. The KAIST dataset [33] is similar to CAVE dataset, but the spatial resolution is much higher than both the CAVE and ICVL datasets. It contains 32 images of size 2704×3376, each of which also consists of 31 bands ranging from 420 nm to 720 nm. In addition, all the KAIST images are normalized by the intensity of the reference white of Spectralon (calibrated 99% reflectance). For these three datasets, we use the integration of the HSIs and the CIE-1964 (as shown in Figure 3) color match function to generate the corresponding RGB images.

In this paper, the proposed SSR-Net is trained based on the PyTorch[2)] framework using the Adam solver [39] for optimization, the beta1 is set as 0.9, the beta2 is set as 0.999. The weight decay is used to reduce the over-fitting problem. The learning rate is decayed exponentially from 0.001 to 0.0001. The training stops when no notable decay of training loss is observed. During training, we import a fixed-size area from the original image into the network of size 64×64. In order to verify the quality of the reconstructed HSIs, several quantitative metrics are adopted, including root mean square error (RMSE), PSNR, and structural similarity index measurement (SSIM).

3.1.2 *Coarse-to-fine based training*

In the proposed coarse-to-fine based strategy, the BP-Net is trained first for coarse estimation. After the BP-Net is well-trained, the RF-Net is then trained by fixing parameters of the BP-Net for fine estimation. In order to demonstrate the superiority of such coarse-to-fine based strategy, an experiment over ICVL dataset is carried out to compare such coarse-to-fine based training strategy with the overall training strategy that training all the parts of the SSR-Net simultaneously. The experimental results of different training strategies for the proposed SSR-Net is shown in Table 3. Obviously, the proposed coarse-to-fine based training strategy outperforms the overall training strategy.

Moreover, experiments on these three datasets are first carried out to demonstrate the superiority of adding RF-Net in the proposed SSR-Net. Comparison between the proposed SSR-Net without RF-Net (actually just BP-Net) and with RF-Net is carried out. The experimental results by two-folded division are listed in Table 4, in which the mean value of PSNR is adopted for quantitative evaluation. It is observed that, about 2 dB improvement can be obtained by adding RF-Net, demonstrating that the RF-Net can clearly alleviate the spectral distortion.

1) http://www1.cs.columbia.edu/CAVE/projects/gap_camera/.
2) http://pytorch.org/.

Table 4 Reconstruction performance of the proposed SSR-Net with and without RF-Net in terms of mean PSNR in dB

	ICVL		CAVE		KAIST	
	Fold 0	Fold 1	Fold 0	Fold 1	Fold 0	Fold 1
BP-Net	40.50	43.76	31.52	30.53	30.49	30.80
SSR-Net	44.66	45.01	33.91	32.36	31.82	32.01

Table 5 Quantitative evaluation results on the three datasets

	Method	RMSE			PSNR (dB)			SSIM			SAM		
		Max	Mean	Std	Min	Mean	Std	Min	Mean	Std	Max	Mean	Std
ICVL	K-SVD	0.0404	0.0136	0.0056	27.88	37.90	**3.08**	0.8804	0.9571	0.0194	0.2038	0.1030	0.0293
	RBF	0.0432	0.0097	0.0064	27.29	41.65	4.65	0.9130	0.9858	0.0112	0.1187	0.0485	0.0163
	HSCNN	0.0296	0.0072	0.0047	30.57	44.03	4.35	0.9653	0.9914	0.0063	0.1062	0.0397	0.0137
	3DCNN	0.0303	0.0069	0.0043	30.38	44.35	4.26	0.9685	0.9918	0.0051	0.1089	0.0375	0.0135
	SSR-Net	**0.0276**	**0.0063**	**0.0040**	**31.17**	**45.12**	4.26	**0.9701**	**0.9929**	**0.0047**	**0.1033**	**0.0337**	**0.0129**
CAVE	K-SVD	0.1066	0.0445	0.0268	19.44	28.33	4.81	0.7481	0.9033	0.0630	0.5360	0.4216	0.0702
	RBF	0.1137	0.0373	0.0300	18.89	30.61	5.78	0.7536	0.9331	0.0585	0.4410	0.2660	0.0782
	HSCNN	**0.0519**	0.0226	**0.0115**	**25.69**	33.86	**4.19**	0.8815	0.9556	0.0268	**0.3431**	0.2017	**0.0525**
	3DCNN	0.0537	0.0226	0.0122	25.40	34.00	4.44	**0.9042**	0.9595	0.0260	0.3621	0.2093	0.0590
	SSR-Net	0.0522	**0.0223**	0.0120	25.65	**34.10**	4.45	0.9041	**0.9599**	**0.0259**	0.3445	**0.2013**	0.0540
KAIST	K-SVD	0.0507	0.0330	0.0080	25.91	29.86	**1.99**	0.8078	0.9075	0.0325	0.6336	0.5173	0.0026
	RBF	0.0392	0.0231	**0.0064**	28.12	33.02	2.23	**0.9008**	0.9467	**0.0184**	0.5280	0.2930	0.0954
	HSCNN	0.0343	0.0191	0.0071	29.30	34.85	2.93	0.8815	**0.9556**	0.0268	0.5839	0.2741	**0.1389**
	3DCNN	0.0362	0.0216	0.0068	28.80	33.68	2.60	0.8494	0.9338	0.0277	0.6824	**0.2193**	0.1628
	SSR-Net	**0.0340**	**0.0186**	0.0072	**29.38**	**35.12**	3.02	0.8607	0.9475	0.0273	**0.5372**	0.2236	0.1534

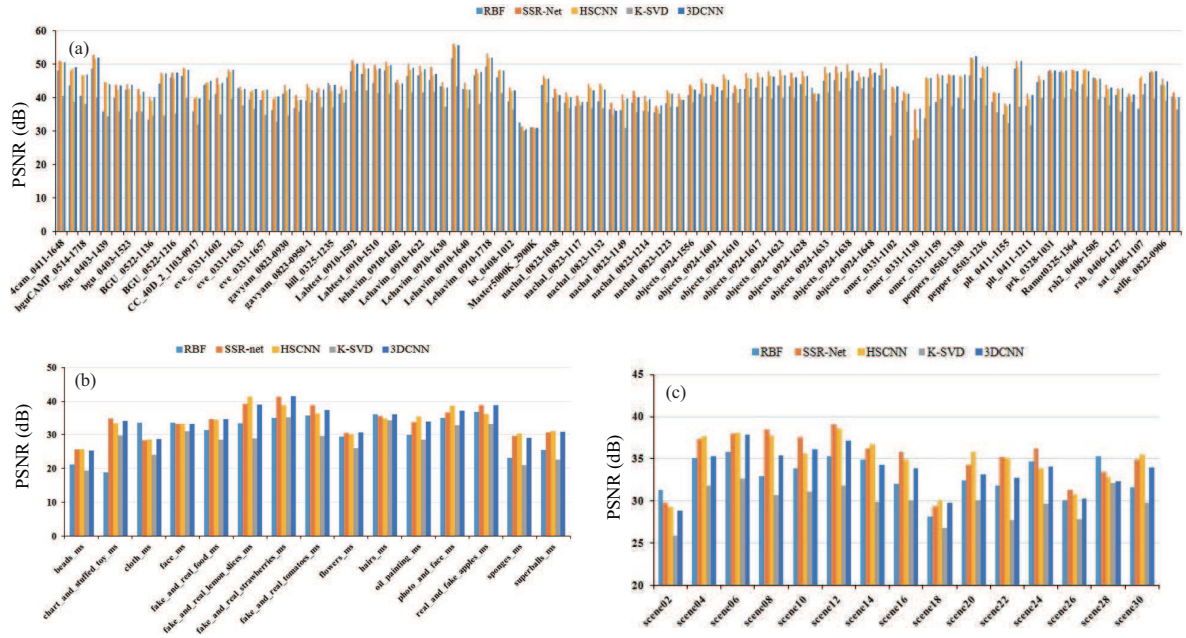

Figure 4 (Color online) Individual results of PSNR over these three datasets. (a) ICVL dataset, (b) CAVE dataset, and (c) KAIST dataset.

3.1.3 *Comparison with state-of-the-art*

In order to demonstrate the effectiveness of the proposed SSR-Net for spectral SR, several state-of-the-art methods, including K-SVD [21], RBF Interpolation [20], HSCNN [24], and 3DCNN in [27], are adopted for comparison. For all of these algorithms, we randomly divide all the considered datasets into two parts: the training set and the testing set, each of which contains half the number of images. The experimental results of all these algorithms are listed in Table 5. For the four quantitative metrics, not only the mean and standard variation of individual results, but also their worst values, i.e., largest RMSE,

Figure 5 (Color online) Sample results of band images reconstructed by different algorithms from 'bgu_0403-1439' in ICVL dataset.

SAM and smallest PSNR, SSIM for all images in the dataset, are used to evaluate the performance of HSI reconstruction. It is observed that the proposed SSR-Net achieves the best performance over all considered algorithms in terms of average RMSE and PSNR on these three datasets, and achieves the best or nearly best performance in terms of average SSIM. Moreover, the performance of the proposed SSR-Net over all the individual results in the three datasets does not vary much in terms of the standard variation. Even when the worse measurements are considered, the proposed SSR-Net also achieves or approaches the best value over all the compared algorithms. Figure 4 further lists the individual results of these three datasets in terms of PSNR. It is also confirmed that the proposed SSR-Net reconstructs HSI with very high accuracy for each individual image.

Three sample results reconstructed by different algorithms from different datasets are also selected and listed in Figures 5–7, respectively. It is also confirmed that our proposed SSR-Net can reconstruct HSIs from RGB input with high quality. Figure 8 further lists the recovered spectra by these algorithms and their corresponding ground-truth spectra from three testing images of different datasets. In each image, three pixels are selected and listed in one row. It is observed that the spectra reconstructed by the proposed SSR-Net are more similar to the ground-truth spectra than the other three algorithms.

Different numbers of training images are also tested on the ICVL dataset, including 5, 10, 15, 20, and 25 images, respectively. All the rest images are used for testing. The experimental results of the proposed SSR-Net with different numbers of training samples are listed in Figure 9. It is observed that the performance of the proposed SSR-Net steadily increases when more samples are used for training.

3.2 Experiments on satellite images for classification

In this subsection, the advantage of spectral SR for the classification task is analyzed by applying the proposed SSR-Net to images acquired by the AVIRIS sensor. Two well-known hyperspectral datasets, i.e., the Indian Pines dataset and the Salinas Valley dataset, are used for evaluation. Figure 10 provides

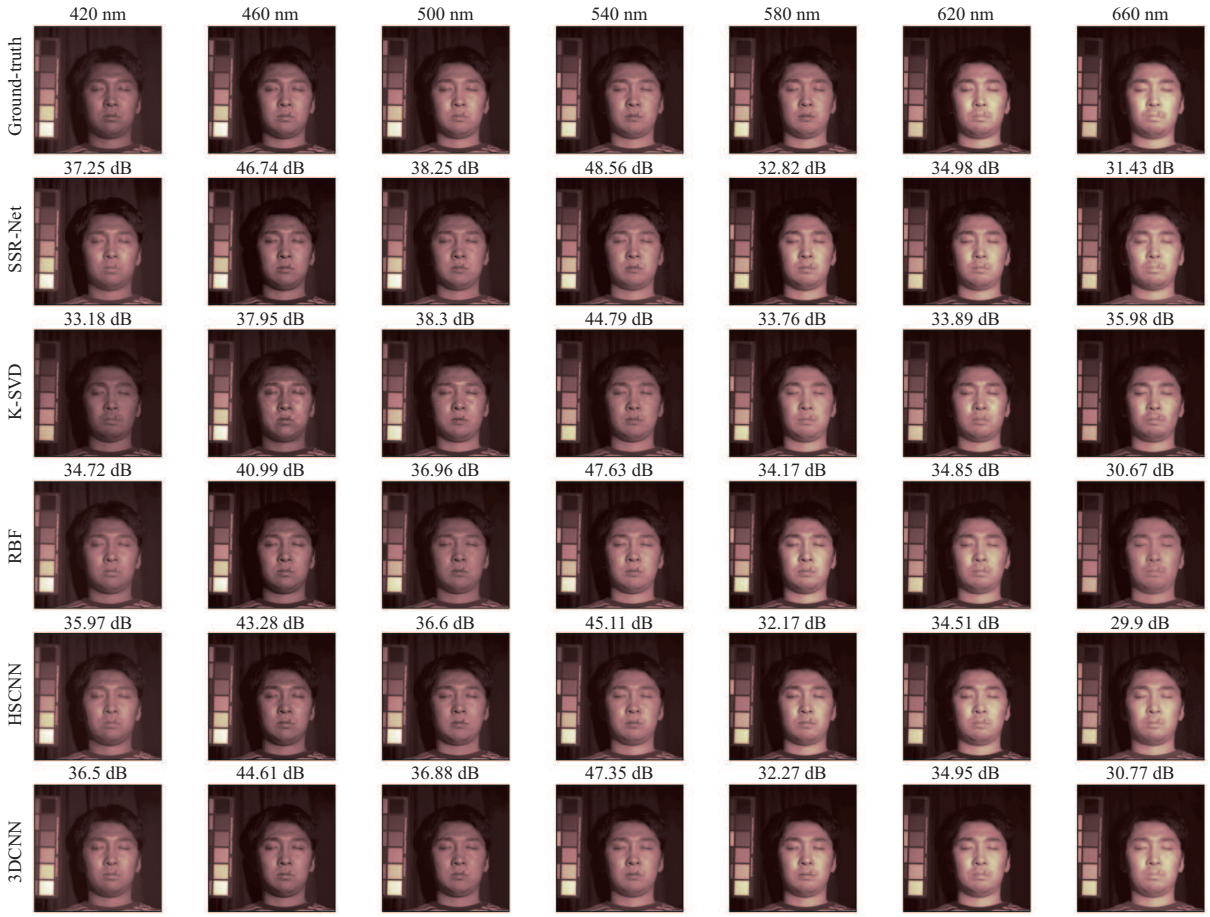

Figure 6 (Color online) Sample results of band images reconstructed by different algorithms from 'face_ms' in CAVE dataset.

Table 6 The classification performance over the Salinas Valley dataset

		OA (%)						AA (%)					
		SSR-Net	3DCNN	HSCNN	RBF	KSVD	RGB	SSR-Net	3DCNN	HSCNN	RBF	KSVD	RGB
SVM		**87.52**	84.74	85.22	78.14	82.35	81.77	**93.24**	91.25	91.53	83.97	89.43	83.70
SVM-CK	3×3	**90.03**	86.81	86.94	82.69	86.85	84.85	**95.23**	92.74	93.06	88.42	93.07	91.69
	5×5	**92.70**	88.46	89.67	86.62	89.11	86.96	**96.69**	93.92	94.34	90.96	94.34	93.25

Table 7 The classification performance over the Indian Pines dataset

		OA (%)						AA (%)					
		SSR-Net	3DCNN	HSCNN	RBF	KSVD	RGB	SSR-Net	3DCNN	HSCNN	RBF	KSVD	RGB
SVM		**63.95**	53.62	58.84	46.57	48.47	53.12	**62.93**	58.60	52.99	53.35	46.49	40.39
SVM-CK	3×3	**73.81**	65.36	72.79	52.90	59.49	66.81	**73.63**	67.56	69.13	61.14	58.79	63.64
	5×5	**79.99**	72.44	77.56	58.59	66.88	72.95	**81.76**	71.94	76.14	67.92	65.67	67.92

the pseudo color image and the ground-truth for classification for these two datasets, respectively. The AVIRIS sensor acquires 224 spectral bands ranging from 0.4 μm to 2.5 μm with a spatial resolution of about 20 m. Owing to its high spectral coverage, AVIRIS scenes have been widely utilized in the remote sensing community for classification. To be consistent with previous experiments, 31 bands covering a range from 400 nm to 700 nm are selected in this experiment. Specifically, these 31 bands correspond to band images from the 6th to the 36th band.

In this experiment, the RGB images are also simulated according to the CIE-1964 color matching functions. Then, these simulated RGB images are used to reconstruct the 31 selected bands of HSI. Each image is divided horizontally into two parts, one for training and the other for validation. Small patches with the size of 24×24 are clipped as input for more training samples. The batch norm method [40] is

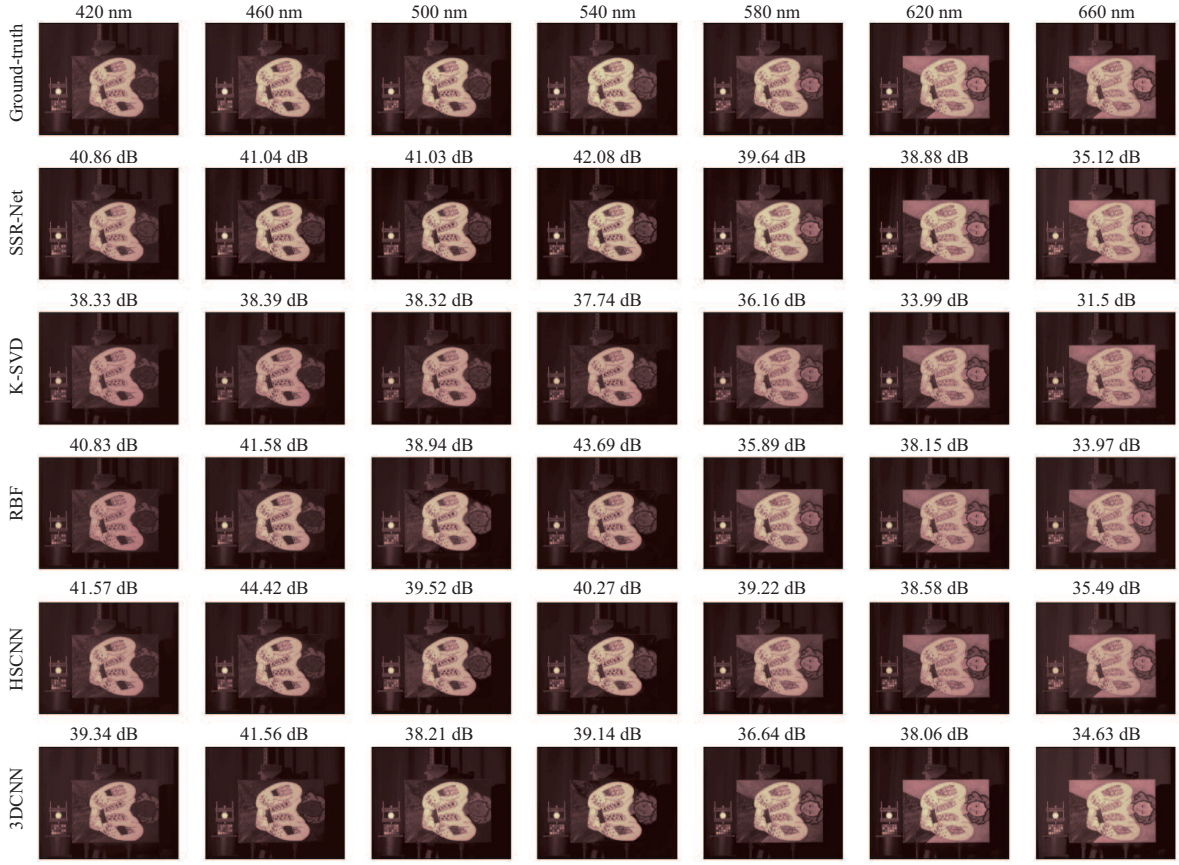

Figure 7 (Color online) Sample results of band images reconstructed by different algorithms from 'scene28_reflectance' in KAIST dataset.

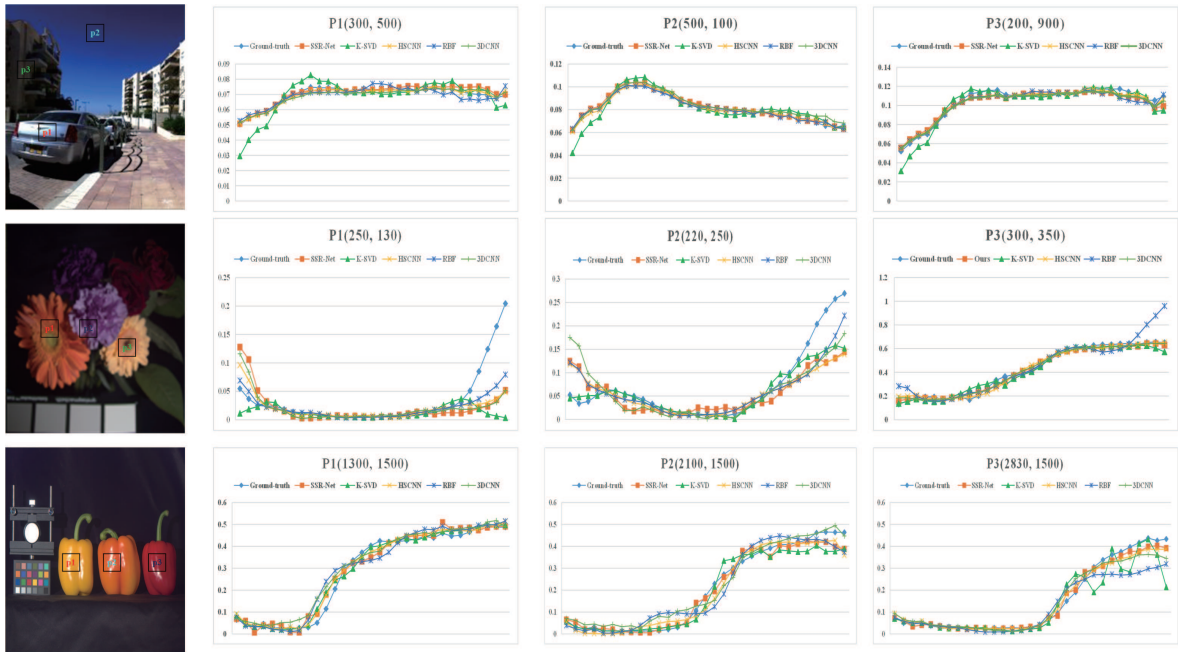

Figure 8 (Color online) Sample results of reconstructed pixels from three datasets. From top to bottom: 'Ramot0325-1364' from ICVL dataset, coordinates: (300, 500), (500, 100), (200, 900); 'flowers_ms' from CAVE dataset, coordinates: (250, 130), (220, 250), (300, 350); 'scene30_reflectance' from KAIST dataset, coordinates: (1300, 1500), (2100, 1500), (2830, 1500).

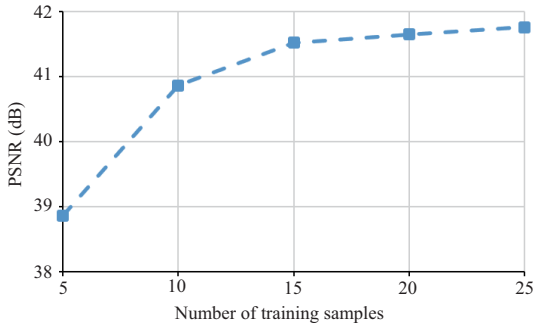

Figure 9 (Color online) Experimental results by adopting different numbers of training samples for the ICVL dataset.

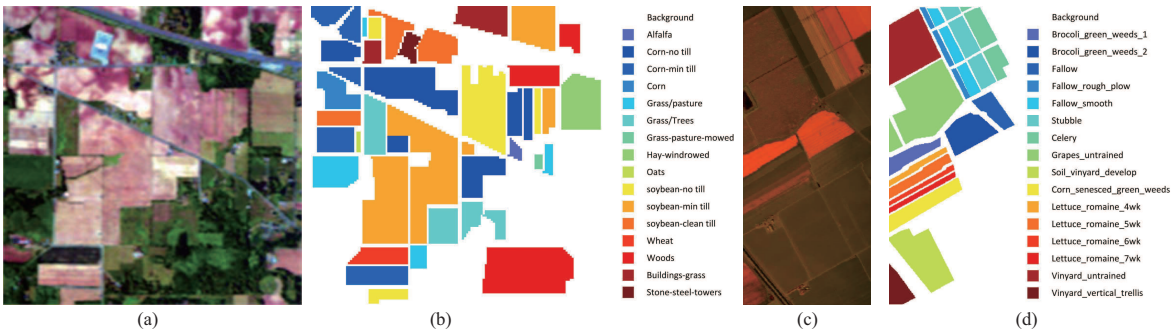

Figure 10 (Color online) (a) A pseudo color image of the Indian Pines dataset. (b) The ground-truth map of the Indian Pines dataset. (c) A pseudo color image of the Salinas Valley dataset. (d) The ground-truth map of the Salinas Valley dataset.

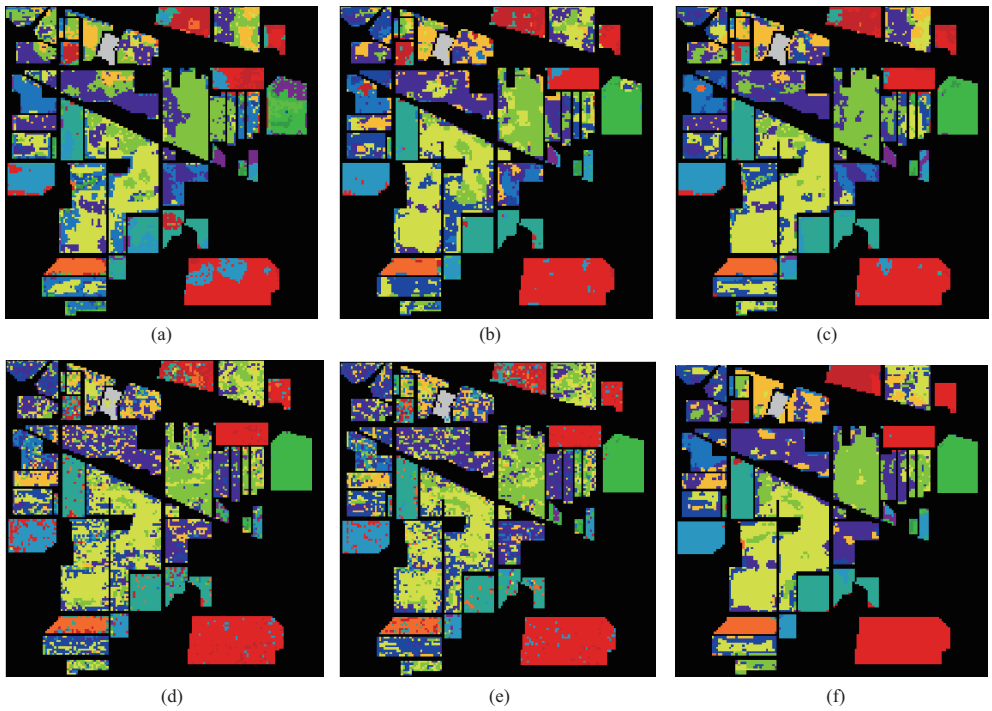

Figure 11 (Color online) Classification maps of different algorithms on the Indian Pines dataset. (a) RGB image; (b) HSI reconstructed by K-SVD [21]; (c) HSI reconstructed by RBF [20]; (d) HSI reconstructed by HSCNN [24]; (e) HSI reconstructed by 3DCNN [27]; and (f) HSI reconstructed by the proposed SSR-Net.

used to expedite the training. For comparison, the K-SVD [21], RBF Interpolation [20], HSCNN [24], and 3DCNN in [27] are also adopted.

After spectral SR, classification is performed on RGB images and their corresponding reconstructed HSIs. Here, two popular classifiers are employed, i.e., support vector machine (SVM) and SVM-CK [41].

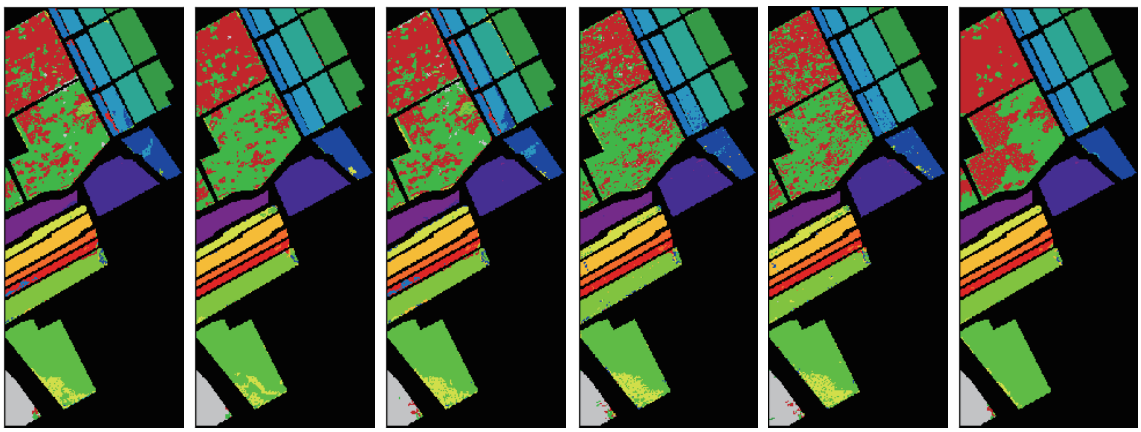

Figure 12 (Color online) Classification maps of different algorithms on the Salinas Valley dataset. (a) RGB image; (b) HSI reconstructed by K-SVD [21]; (c) HSI reconstructed by RBF [20]; (d) HSI reconstructed by HSCNN [24]; (e) HSI reconstructed by 3DCNN [27]; and (f) HSI reconstructed by the proposed SSR-Net.

In each dataset, 10% of the pixels in each category were used to train the classifier. As shown in Tables 6 and 7, the classification performance of RGB images was improved by reconstructing the HSIs, which indicates that spectral SR is very relative to improving applications performance, e.g., classification tasks. In addition, the proposed SSR-Net clearly outperformed the other three algorithms relative to reconstructing HSIs from RGB images. This conclusion is confirmed by the classification maps obtained by different algorithms on the Indian Pines (Figure 11) and Salinas Valley (Figure 12) datasets.

4 Conclusion and future work

In this paper, we have explored the intrinsic relationship between RGB images and HSIs using a CNN, and we have proposed the SSR-Net for spectral SR of RGB images. In addition to an implementing BP-Net to predict band images directly from RGB inputs, the proposed SSR-Net employs RF-Net to further improve spectral fidelity. As a result, HSIs can be reconstructed directly from RGB images without any other priors. The experimental results obtained on three benchmark datasets demonstrate that the proposed SSR-Net outperforms state-of-the-art methods. In addition, the experimental results obtained on satellite images demonstrate that classification performance can be improved by spectral SR using the proposed SSR-Net.

Although the superior performance was obtained by the proposed SSR-Net, many training samples are required in such deep learning-based techniques to reconstruct HSIs using RGB inputs. Thus, in the future, it would be valuable to further investigate few-shot learning-based strategies to learn such mappings. In addition, the transferring ability to RGB images with different contents with training images will also be considered.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant No. 61671383), Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University (Grant No. CX2020020), and Hong Kong RGC (Grant No. 9042820 (CityU 11219019)).

References

1 Li J, Marpu P R, Plaza A, et al. Generalized composite kernel framework for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2013, 51: 4816–4829
2 Li W, Du Q, Zhang B. Combined sparse and collaborative representation for hyperspectral target detection. Pattern Recogn, 2015, 48: 3904–3916
3 Huang X, Zhang L. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. IEEE Trans Geosci Remote Sens, 2013, 51: 257–272
4 Ma M, Mei S, Wan S, et al. Video summarization via block sparse dictionary selection. Neurocomputing, 2020, 378: 197–209
5 Zhang Z J, Pang Y W. CGNet: cross-guidance network for semantic segmentation. Sci China Inf Sci, 2020, 63: 120104
6 Ma S, Pang Y W, Pan J, et al. Preserving details in semantics-aware context for scene parsing. Sci China Inf Sci, 2020, 63: 120106
7 Xie J, Pang Y W, Cholakkal H, et al. PSC-Net: learning part spatial co-occurrence for occluded pedestrian detection. Sci China Inf Sci, 2021, 64: 120103
8 Cao J, Pang Y, Li X. Learning multilayer channel features for pedestrian detection. IEEE Trans Image Process, 2017, 26: 3210–3220

9 Alparone L, Wald L, Chanussot J, et al. Comparison of pansharpening algorithms: outcome of the 2006 GRS-S data-fusion contest. IEEE Trans Geosci Remote Sens, 2007, 45: 3012–3021
10 Bendoumi M A, He M Y, Mei S H. Hyperspectral image resolution enhancement using high-resolution multispectral image based on spectral unmixing. IEEE Trans Geosci Remote Sens, 2014, 52: 6574–6583
11 Zhang Y. Spatial resolution enhancement of hyperspectral image based on the combination of spectral mixing model and observation model. In: Proceedings of SPIE, 2014. 9244: 201–204
12 Li X, Ling F, Foody G M, et al. Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps. Remote Sens Environ, 2017, 196: 293–311
13 Fu Y, Zheng Y, Huang H, et al. Hyperspectral image super-resolution with a Mosaic RGB image. IEEE Trans Image Process, 2018, 27: 5539–5552
14 Zhang L, Wei W, Zhang Y, et al. Cluster sparsity field: an internal hyperspectral imagery prior for reconstruction. Int J Comput Vis, 2018, 126: 797–821
15 Liebel L, Körner M. Single-image super resolution for multispectral remote sensing data using convolutional neural networks. Int Arch Photogramm Remote Sens Spatial Inf Sci, 2016, XLI-B3: 883–890
16 Li Y, Hu J, Zhao X, et al. Hyperspectral image super-resolution using deep convolutional neural network. Neurocomputing, 2017, 266: 29–41
17 Hu J, Li Y, Xie W. Hyperspectral image super-resolution by spectral difference learning and spatial error correction. IEEE Geosci Remote Sens Lett, 2017, 14: 1825–1829
18 Mei S, Yuan X, Ji J, et al. Hyperspectral image spatial super-resolution via 3D full convolutional neural network. Remote Sens, 2017, 9: 1139
19 Mei S, Jiang R, Li X, et al. Spatial and spectral joint super-resolution using convolutional neural network. IEEE Trans Geosci Remote Sens, 2020. doi: 10.1109/TGRS.2020.2964288
20 Nguyen R M H, Prasad D K, Brown M S. Training-based spectral reconstruction from a single RGB image. In: Proceedings of European Conference on Computer Vision, 2014. 186–201
21 Arad B, Ben-Shahar O. Sparse recovery of hyperspectral signal from natural RGB images. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016. 19–34
22 Yi C, Zhao Y Q, Chan J C W. Spectral super-resolution for multispectral image based on spectral improvement strategy and spatial preservation strategy. IEEE Trans Geosci Remote Sens, 2019, 57: 9010–9024
23 Jia Y, Zheng Y, Gu L, et al. From RGB to spectrum for natural scenes via manifold-based mapping. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV), 2017. 4715–4723
24 Arad B, Ben-Shahar O, Timofte R, et al. NTIRE 2018 challenge on spectral reconstruction from RGB images. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Los Alamitos, 2018
25 Can Y B, Timofte R. An efficient CNN for spectral reconstruction from RGB images. 2018. ArXiv: 1804.04647
26 Han X, Yu J, Xue J, et al. Spectral super-resolution for RGB images using class-based BP neural networks. In: Proceedings of 2018 Digital Image Computing: Techniques and Applications (DICTA), 2018. 1–7
27 Koundinya S, Sharma H, Sharma M, et al. 2D-3D CNN based architectures for spectral reconstruction from RGB images. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018
28 Zhang L, Zhang L, Du B. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci Remote Sens Mag, 2016, 4: 22–40
29 Mei S, Ji J, Hou J, et al. Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks. IEEE Trans Geosci Remote Sens, 2017, 55: 4520–4533
30 Yuan Q, Zhang Q, Li J, et al. Hyperspectral image denoising employing a spatial-spectral deep residual convolutional neural network. IEEE Trans Geosci Remote Sens, 2019, 57: 1205–1218
31 Zhang M, Li W, Du Q. Diverse region-based CNN for hyperspectral image classification. IEEE Trans Image Process, 2018, 27: 2623–2634
32 Yasuma F, Mitsunaga T, Iso D, et al. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. IEEE Trans Image Process, 2010, 19: 2241–2253
33 Choi I, Jeon D S, Nam G, et al. High-quality hyperspectral reconstruction using a spectral prior. ACM Trans Graph, 2017, 36: 1–13
34 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv: 1409.1556
35 Chen L, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. 2017. ArXiv: 1706.05587
36 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proceedings of International Conference on Learning Representations, 2016
37 Zhao H, Gallo O, Frosio I, et al. Loss functions for image restoration with neural networks. IEEE Trans Comput Imag, 2017, 3: 47–57
38 Saxe A M, Mcclelland J L, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In: Proceedings of International Conference on Learning Representations, 2014. 1–22
39 Kingma D P, Ba J. Adam: a method for stochastic optimization. 2014. ArXiv:1412.6980
40 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. ArXiv: 1502.03167
41 Camps-Valls G, Gomez-Chova L, Munoz-Mari J, et al. Composite kernels for hyperspectral image classification. IEEE Geosci Remote Sens Lett, 2006, 3: 93–97