Neural Networks 146 (2022) 107–119

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Dilated projection correction network based on autoencoder for hyperspectral image super-resolution

Xinya Wang [a], Jiayi Ma [a,*], Junjun Jiang [b], Xiao-Ping Zhang [c]

[a] Electronic Information School, Wuhan University, Wuhan, 430072, China
[b] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[c] Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada

ARTICLE INFO

Article history:
Received 16 December 2020
Received in revised form 7 September 2021
Accepted 11 November 2021
Available online 17 November 2021

Keywords:
Hyperspectral image
Super-resolution
Deep learning
Autoencoder

ABSTRACT

This paper focuses on improving the spatial resolution of the hyperspectral image (HSI) by taking the prior information into consideration. In recent years, single HSI super-resolution methods based on deep learning have achieved good performance. However, most of them only simply apply general image super-resolution deep networks to hyperspectral data, thus ignoring some specific characteristics of hyperspectral data itself. In order to make full use of spectral information of the HSI, we transform the HSI SR problem from the image domain into the abundance domain by the dilated projection correction network with an autoencoder, termed as *aeDPCN*. In particular, we first encode the low-resolution HSI to abundance representation and preserve the spectral information in the decoder network, which could largely reduce the computational complexity. Then, to enhance the spatial resolution of the abundance embedding, we super-resolve the embedding in a coarse-to-fine manner by the dilated projection correction network where the back-projection strategy is introduced to further eliminate spectral distortion. Finally, the predictive images are derived by the same decoder, which increases the stability of our method, even at a large upscaling factor. Extensive experiments on real hyperspectral image scenes demonstrate the superiority of our method over the state-of-the-art, in terms of accuracy and efficiency.

1. Introduction

Hyperspectral imaging collects spectral information of the same scene over a large number of continuous and narrow spectral bands, ranging from ultraviolet to infrared wavelength. Thus, the hyperspectral image (HSI) acquires both spatial relationships and reflectance nature of different objects with a high spectral resolution. As a 3D data cube, each pixel of HSI represents a spectral curve of the related material, which can be utilized to distinguish the objects in the image scene, especially the earth objects. Because of this property, HSI has proven to be useful for many occasions, such as land surface classification (Liu, Gu, Chanussot, & Dalla Mura, 2017; Wu & Prasad, 2017), anomaly detection (Du & Zhang, 2014; Xie et al., 2020), environmental monitoring (Plaza, Du, Bioucas-Dias, Jia, & Kruse, 2011), and so on. However, due to the inevitable trade-off between spatial resolution, spectral resolution, and signal-to-noise ratio, HSI usually has a low spatial resolution, limiting the range of potential applications (Yokoya, Grohnfeldt, & Chanussot, 2017). For this reason, it is essential to develop a software technique that can improve the spatial resolution of HSI without losing detailed spectral information (Liu, Wen, Fan, Loy, & Huang, 2018; Wen, Kamilov, Liu, Mansour, & Boufounos, 2018). This technique is known as super-resolution (SR), which is a hot topic in computer vision.

SR is a classical method to reconstruct high-resolution (HR) image through one or more corresponding low-resolution (LR) images. It is a post-processing method to improve spatial resolution without modifying the hardware. Most existing SR methods designed for HSI are fusion-based methods (Akhtar, Shafait, & Mian, 2015; Dong et al., 2016; Qu, Qi, & Kwan, 2018; Simões, Bioucas-Dias, Almeida, & Chanussot, 2014; Veganzones et al., 2015; Wei, Bioucas-Dias, Dobigeon, & Tourneret, 2015), which recover the HR HSI by fusing the LR HSI with corresponding auxiliary images of the same scene. These additional observations usually have high spatial resolution with a low spectral resolution, including the multi-spectral image, the panchromatic image, and the RGB image. Nevertheless, due to the particularity of the hyperspectral imaging scene, these auxiliary images captured at the same scene as HSI are often scarce even unavailable. As a result, the absence of auxiliary images hinders the practical

* Corresponding author.
E-mail addresses: wangxinya@whu.edu.cn (X. Wang), jyma2010@gmail.com (J. Ma), junjun0595@163.com (J. Jiang), xzhang@ee.ryerson.ca (X.-P. Zhang).

https://doi.org/10.1016/j.neunet.2021.11.014
0893-6080/

application of the fusion-based SR methods. Another obstacle to some fusion-based methods is the premise that the auxiliary observation has been fully registered with HSI, which is also a challenging task. Therefore, it is valuable to explore the HSI SR method based on a single input. The single-image SR method produces HR HSI from a single LR HSI without any additional information.

In the past decades, many single-image-based methods have been proposed to solve the HSI SR problem (Jiang, Sun, Liu, & Ma, 2020; Wang, Ma, & Jiang, 2021). Some traditional approaches typically use a prior (which is sophisticatedly designed by hand) such as self-similarity, sparsity, or low rank of the HSI to regularize the super-resolution reconstruction process (He, Zhou, Wang, Cao and Han, 2016; Huang, Yu, & Sun, 2014). Recently, due to the strong ability to extract image features, the convolutional neural network (CNN) has been successfully applied to reconstruct HR HSI in an end-to-end manner. Although all of them are proven to be effective in this task, there still exist some shortcomings. Specifically, the former's shallow heuristic models have limited expressive power. Thus, they usually fail to appropriately recover the complex image details. Taking the whole data with hundreds of spectral bands as the input, the CNN-based methods either regard the HSI as ordinary 3D data cube, ignoring the spectral characteristics of the HSI, or have high computational complexity.

To address the above-mentioned challenges, we consider the basic spectrum of the HSI as prior information to solve the HSI SR problem in the abundance domain, which is implemented by the dilated projection correction network with autoencoder, termed as *aeDPCN*. In our method, we transform the HSI SR problem into the abundance domain by autoencoder based on matrix factorization. According to the hypothesis of the linear spectral mixture model, nonnegative matrix factorization (NMF) as a useful unmixing method can identify endmember spectra and estimate the corresponding abundances of HSI. In particular, describing the same imaging scene, LR and HR HSIs should have the same endmember. Due to this observation, we transform the nonnegative hyperspectral data through an autoencoder network which plays the role of decomposition and reconstruction. Concretely, we embed the nonnegative and sum-to-one constraints into the encoder to generate the abundance representation while the decoder preserves the basic spectral information. To improve the spatial resolution, the dilated projection correction network (DPCN) is adopted to super-resolve the abundance embedding that is much smaller than the original HSI. In this network, we progressively reconstruct the HR embedding and use the hybrid dilated convolution to extract informative features with a large reception field. Sharing the same decoder, the generated HR representation is required to follow a similar pattern as the LR representation. Therefore, we design the back-projection module to constrain the generated abundance for correcting the spectral distortion. Compared with other CNN-based methods, we perform super-resolution in the abundance domain, which could greatly reduce the computational complexity. Besides, the spectral prior has been preserved in the decoder, which makes our method more accurate for reconstruction without much spectral loss, even at a large scale factor.

As shown in Fig. 1, spectral decomposition is also a general procedure for fusion-based methods (Lanaras, Baltsavias, & Schindler, 2015; Qu et al., 2018; Yokoya, Yairi, & Iwasaki, 2011). In this procedure, through model-based methods or learning-based methods, the endmember spectrum S and the HR abundance A_h are extracted from the LR HSI and the HR auxiliary observation, respectively. On the contrary, our proposed aeDPCN redefines the SR problem in the abundance domain. In particular, the abundance representation is regarded as the embedding to improve the spatial resolution. Due to the absence of auxiliary observations for single HSI SR, the dilated projection correction network attempts to learn a mapping function between the LR abundance $\mathcal{A}_l$ and the corresponding HR one $\mathcal{A}_h$, which has been rarely studied in the HSI SR problem.

In summary, our major contribution is three-fold. First, we solve the SR problem in the abundance domain, which is the first attempt in the single HSI SR task. Specifically, we transform the single HSI SR problem into the abundance domain and unify the matrix transformation and SR in an end-to-end manner. In this way, the computational overhead, the critical issue in HSI SR problem, could be greatly reduced and spectral information can be better preserved. Second, we deploy the hybrid dilated convolution in limited layers to acquire deep spatial features for super-resolving the abundance representation. As a result, our aeDPCN is a light-weight model with comparable performance, which is demonstrated in the extensive experiments. Last but not least, when improving the spatial resolution, to ensure that the HR and LR abundance representations have similar patterns, a back-projection correction strategy is integrated to reduce the spectral distortion.

2. Related work

2.1. Deep neural network for SR

With the expansion of image databases and the progress of computational technology, learning-based methods have made great achievements in high-level computer vision tasks, such as image classification, objection detection, and scene segmentation. Similarly, to low-level problems such as super-resolution, learning-based methods also have achieved significantly excellent results in nature images mainly through the deep neural network (DNN). Due to its powerful learning ability, DNN is designed to learn a mapping function between LR and HR image pairs in an end-to-end manner. Dong, Loy, He, and Tang (2015) first applied three convolutional layers to resolve the SR problem successfully, which has shown great superiority to the traditional SR methods. Subsequently, through skip connection, residual learning (He, Zhang, Ren and Sun, 2016) was introduced in this task to ease the training process of the deep network, such as very deep network for super-resolution (VDSR) (Kim, Kwon Lee, & Mu Lee, 2016a), deeply recursive convolutional network (DRCN) (Kim, Kwon Lee, & Mu Lee, 2016b), and deeply recursive residual network (DRRN) (Tai, Yang and Liu, 2017). Besides, enhanced deep super-resolution network (EDSR) (Lim, Son, Kim, Nah, & Mu Lee, 2017) removed unnecessary parts in the residual modules and won the NTIRE2017 Super-Resolution Challenge. For maximum feature reuse, in Tai, Yang, Liu and Xu (2017), Tong, Li, Liu, and Gao (2017) and Zhang, Tian, Kong, Zhong and Fu (2018), the outputs of convolutional layers were densely connected to batter propagate information. After that, researchers have investigated deeper networks or more sophisticated structures to reconstruct HR images. For instance, the generative adversarial network (GAN) (Goodfellow et al., 2014) was proposed in Ledig et al. (2017) and Wang et al. (2018) for photo-realistic SR, which can alleviate the blurring and over-smoothing artifacts to some extend. To generate multi-scale predictions, a progressive reconstruction was adopted by the Laplacian pyramid SR network (LapSRN) (Lai, Huang, Ahuja, & Yang, 2017) in a coarse-to-fine manner. Similarly, Haris, Shakhnarovich, and Ukita (2018) exploited iterative up-and-down downsampling layers with an error feedback mechanism to construct deep back-projection networks (DBPN), establishing new state-of-the-art results for large scaling factors. From the innovation of classification task, residual channel attention networks (RCAN) (Zhang et al., 2018) utilized attention mechanism in a residual way and stacked up to 400 layers, which achieved the best performance.

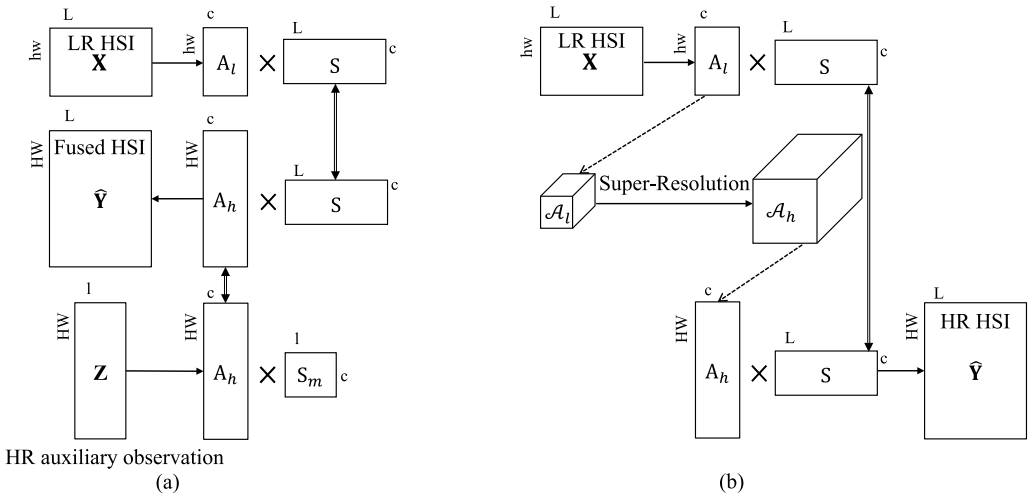

Fig. 1. Difference between the fusion methods based on spectral decomposition and our aeDPCN. (a) General procedure of fusion methods based on spectral decomposition. (b) The procedure of our proposed aeDPCN.

Although the natural image SR problem has been extensively studied and these aforementioned methods have achieved good results on RGB images, most of them cannot be applied for HSI SR directly. On the one hand, in general, these methods are trained to process three-channel or single-channel images. Since the HSI data have hundreds of channels, these CNN-based methods pre-trained on RGB images should be extended in a band-by-band manner to super-resolve single-band image. However, this way could cause spectral distortion because each pixel of the HSI data reflects spectral information of the specific material. On the other hand, if the input and output dimensions of these models are adjusted to the corresponding HSI data dimension, we should train the network from scratch. In order to avoid over-fitting, most very deep networks need a lot of data to drive. There are not enough hyperspectral data to meet this requirement. Therefore, it is necessary to design lightweight methods specifically for HSI.

2.2. CNN for HSI SR

Recently, several CNN-based methods have been proposed for HSI SR. To alleviate spectral distortion, a spectral difference convolutional neural network (SECNN) was proposed with the combination of a spatial-error correction (SEC) model (Hu, Li, & Xie, 2017; Xie, Li, Hu, & Chen, 2018), which achieved both spatial information enhancement and spectral information preservation. Similarly, Li, Hu, Zhao, Xie, and Li (2017) combined a spatial constraint (SCT) strategy with SDCNN model to make the LR HSI generated by the reconstructed HSI spatially close to the input LR HSI. To further exploit spatial–spectral information, Hu, Zhao, and Li (2019) presented an intra-fusion operation based on a deep information distillation network. Subsequently, Yuan, Zheng, and Lu (2017) exploited transfer learning with collaborative non-negative matrix factorization (CNMF) to enforce collaborations between the LR and HR HSIs. Sharing the same idea of matrix factorization, deep feature matrix factorization (DFMF) (Xie, Jia, Li and Lei, 2019) blended feature matrix extracted by a DNN with NMF strategy for super-resolving real-scene HSI. In general, the SR problem in these methods was solved in a multi-step optimization strategy. Firstly, due to the redundancy of hyperspectral bands, the band selection is conducted on the original HSI data. Then, in some way, the key bands would be super-resolved by the CNN to improve their spatial resolution. Lastly, the post-processing strategy utilizes the HR key bands to restore the HR HSI and corrects its spectral information. Although these methods have achieved good results, they rely heavily on manual processing and consume a lot of time. Hence, some methods implemented in an end-to-end manner for HSI SR have been proposed. Mei et al. (2017) designed a three-dimensional full CNN (3D-FCNN) method to exploit both the spatial context of neighboring pixels and spectral correlation of neighboring bands. Although 3D convolution could capture both spatial and spectral correlation, the computational complexity is comparatively high, especially at a large input. Due to spectral disorder caused by normal 2D convolution, grouped deep recursive residual network (GDDRN) (Li, Zhang, Dingl, Wei, & Zhang, 2018) embedded a grouped recursive module into the global residual structure with a joint loss to reduce both the numerical error and spectral distortion. Similarly, separable-spectral and inception network (SSIN) (Zheng et al., 2019) extracted features of each band image independently by separable-spectral convolution and fused them in a coarse-to-fine manner. Recently, a novel mixed convolutional network (MCNet) (Li, Wang, & Li, 2020) is designed to extract the potential features by 2D/3D convolution for mining spatial features of hyperspectral image. However, all of these end-to-end methods conduct on the original HSI data directly and take them as ordinary data cube for resolution lifting, which neglect the spectrum knowledge of HSI. Under the circumstances, they cannot preserve the spectral information well. Therefore, according to the transformation conducted by the autoencoder, we collect the basic spectra in the decoder as prior information and use the dilated back-projection correction network to learn the mapping function between LR and HR abundance maps.

3. Problem formulation

Given the LR HSI, $\mathcal{X} \in \mathbb{R}^{h\times w\times L}$ with L spectral bands, where h and w denote the width and height of each band, the goal of the single-image SR method for HSI is to estimate the corresponding HR HSI, $\hat{\mathcal{Y}} \in \mathbb{R}^{H\times W\times L}$ with the enhancement of spatial resolution at the upscale factor s, in which $H = s \times h$, $W = s \times w$. Hence, $\hat{\mathcal{Y}}$ is expected to as close to the ground-truth HR HSI $\mathcal{Y} \in \mathbb{R}^{H\times W\times L}$ as possible.

In general, the existing CNN-based methods learn the mapping function between LR HSI and HR HSI, which can be defined as:

$$\hat{\mathcal{Y}} = \mathcal{M}(\mathcal{X}; \theta), \tag{1}$$

where $\mathcal{M}$ is the SR model and θ refers to the parameter of $\mathcal{M}$. In this case, these methods handle the original input as ordinary data cube and ignore the prior information of HSI, which would

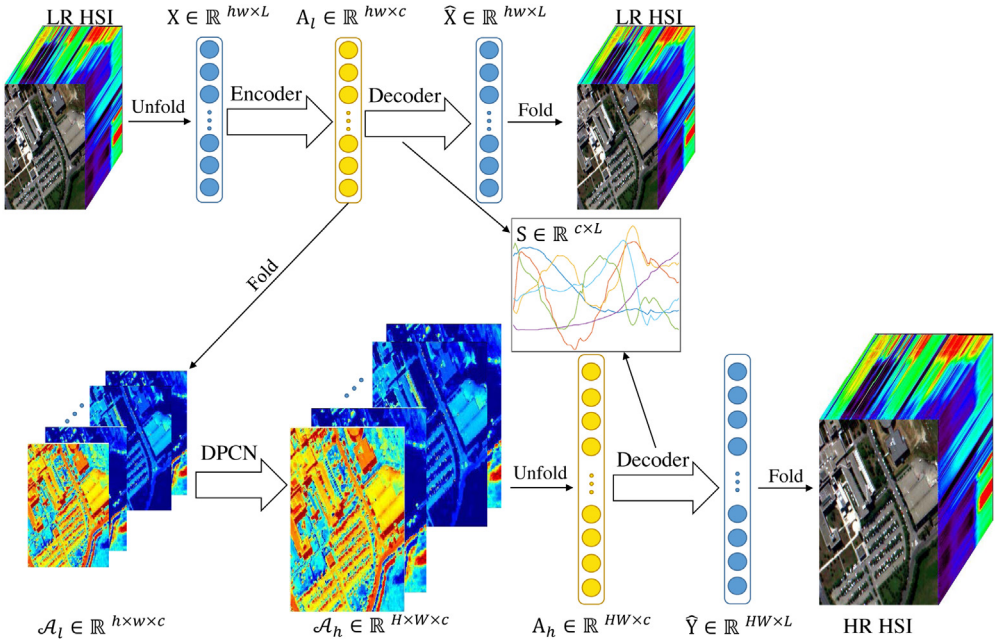

Fig. 2. The whole diagram of our proposed aeDPCN. DPCN represents the dilated back-projection correction network.

cause the loss of spectral information while improving spatial resolution.

To address this issue, we take the endmember spectrum as prior information and redefine the SR problem in the abundance domain. To facilitate the matrix factorization, we unfold the 3*D* data cube into the 2*D* matrix, *i.e.* each row of the 2*D* map denotes the spectral reflectance of a given pixel. The LR HSI data can be rewritten as $X \in \mathbb{R}^{hw\times L}$. Due to its physical effectiveness and mathematical simplicity, a linear spectral mixture model is widely used for unmixing HSI. Thus, we suppose that each row of the unfolded HSI data is a linear combination of c endmember spectra and the LR nonnegative matrix can be factorized into:

$$X = A_l S, \tag{2}$$

where $A_l \in \mathbb{R}^{hw\times c}$ is the abundance matrix with each row vector denoting the abundance fractions of all endmembers at that pixel, $S \in \mathbb{R}^{c\times L}$ represents the endmember spectrum and c is the number of endmembers with $c \ll L$. Since HR and LR observations describe the same scene, the underlying materials (*i.e.*, endmembers) should be the same. Therefore, if we improve the spatial resolution of the abundance matrix, sharing the same endmember spectrum, the corresponding HR HSI $\hat{Y} \in \mathbb{R}^{HW\times L}$ could be reconstructed by:

$$\hat{Y} = A_h S, \tag{3}$$

where $A_h \in \mathbb{R}^{HW\times c}$ is the predicted abundance map of HR HSI.

To this end, our goal is to learn the mapping function between LR and HR abundance matrices. Since the abundance coefficients indicate how the spectral bases are mixed at specific spatial locations, they still preserve the spatial structure of the original HSI. In order to capture its spatial relationship, we fold the abundance matrices as $\mathcal{A}_l \in \mathbb{R}^{h\times w\times c}$ and $\mathcal{A}_h \in \mathbb{R}^{H\times W\times c}$. Thus, the mapping function is:

$$\mathcal{A}_h = \mathcal{M}(\mathcal{A}_l; \theta). \tag{4}$$

In this way, we transform the SR problem into the abundance domain, which reduces the complexity of the model. By taking advantage of the endmember spectrum as the prior information, we would preserve as many spectral details as possible while improving the spatial resolution.

4. Proposed method

Fig. 2 depicts the diagram of our proposed aeDPCN. The 3*D* LR HSI is first unfolded into 2*D* matrices for spectral decomposition, which is implemented by the autoencoder network. In order to transform the SR problem into the abundance domain, the encoder tries to generate the abundance coefficients, satisfying the nonnegative and sum-to-one constraints. Thereby, we can reduce the dimensionality of the data from $\mathbb{R}^{hw\times L}$ to $\mathbb{R}^{hw\times c}$ where $c \ll L$. In addition, inspired by the basic idea of NMF, the decoder network is utilized to represent the endmember matrix for reconstructing HSI, while preserving the spectral information contained in the HSI. Then we consider the abundance representation of LR HSI as an embedding of the HSI. As the spatial structure of the HSI is still preserved in the abundance domain, the HSI embedding is folded and fed into the dilated projection correction network to improve the spatial resolution. The super-resolved embedding gives an estimation of the HR abundance coefficient. Since the HR HSI and LR HSI are sampled from the same scene, the same endmember matrix is multiplied on the abundance representation to produce the HR HSI by sharing the same decoder.

4.1. Autoencoder-based spectral decomposition

Widely used for dimension reduction and representation learning, the autoencoder-based network attempts to approximate an identity map, and accordingly, the output is close to the input. This architecture usually consists of an encoder to map the input data to low-dimensional representations and a decoder for reconstitution, which is suitable to implement the process of Eq. (2) and widely used for the unmixing problem (Palsson, Sigurdsson, Sveinsson, & Ulfarsson, 2018; Qu & Qi, 2018; Su et al., 2019; Xie, Lei, Liu, Li and Jia, 2019). In general, the abundance and the endmember can be generated by this structure or represented by its parameters.

For the sake of the SR problem in Eq. (4), we encode the original input into low-dimensional abundance representation via

$$A_l = f_e(X; \theta_e), \tag{5}$$

Table 1
The detailed structure of the autoencoder network.

	Layers	Activation function	Unit
Encoder	Input layer	LeakyReLU	L
	Hidden layer	LeakyReLU	8c
	Hidden layer	LeakyReLU	4c
	Hidden layer	LeakyReLU	2c
	Hidden layer	ReLU + Normalization	c
Decoder	Output layer	–	L

where f_e denotes the operation of the encoder with the parameter θ_e. The encoder network consists of several fully-connected layers. It is essential that the activation function should be nonlinear, otherwise the encoder would simply perform the dimensional reduction.

Based on the linear spectral mixture model, the abundances represent relative fractions of the spectral bases. Therefore, they are required to satisfy the non-negative constraint and sum-to-one constraint, that are

$$a_{ij} \geq 0, \forall i, j, \tag{6}$$

$$\sum_{j=1}^{c} a_{ij} = 1. \tag{7}$$

For the non-negative constraint, some works have introduced a threshold (Qu & Qi, 2018) or a nonnegative autoencoder (Qu et al., 2018; Su et al., 2019) to enforce the vector to be nonnegative. Besides, the regularization (Qu et al., 2018; Su et al., 2019), or normalization (Palsson et al., 2018; Su et al., 2019) operators have been employed to guarantee the sum-to-one constraint. In order to avoid the complex design of the network, we use the ReLU activation function at the end of the encoder to enforce the output of the encoder to be nonnegative and this nonnegative vector is normalized by the sum of its entries, namely

$$a_{ij} = \frac{a_{ij}}{\sum_{j=1}^{c} a_{ij}}. \tag{8}$$

After acquiring the low-dimensional embedding, the decoder attempts to reconstruct the input spectrum as faithfully as possible by

$$\hat{X} = f_d(A_l, \theta_d), \tag{9}$$

in which f_d denotes the operation of the decoder with the parameter θ_d. Compared to Eq. (2), it is evident that the decoder has to be a linear transformation $f_d : \mathbb{R}^{hw \times c} \rightarrow \mathbb{R}^{hw \times L}$. Consequently, the weight of the decoder is expected as the endmember matrix S carrying spectral information. Thus we have

$$S = \theta_d. \tag{10}$$

To this end, taking the abundance matrix A as input, the decoder network could reconstruct the HSI data, which simulates the Eqs. (2) and (3). The detailed structure of the autoencoder network is reported in Table 1.

Through the autoencoder-based network, we transform the single HSI SR problem into the abundance domain and preserve the spectral prior to HR reconstruction. Since both HR and LR HSIs capture the same imaging scene, it is reasonable to share the same decoder.

4.2. Dilated projection correction network for SR

The objective of this network is to map the LR abundance representation into HR space. Since the spatial relationship of the original data is still preserved in the abundance matrix, we fold it into three-dimensional data to facilitate feature extraction by convolutional layers. As shown in Fig. 3 that depicts the SR network at the scale factor 4, our model has two basic parts: the dilated feature reconstruction module to increase spatial resolution and the back-projection correction module for spectral correction.

4.2.1. Dilated feature reconstruction

Inspired by Lai et al. (2017), we super-resolve the LR embedding in a coarse-to-fine manner. Specifically, the HR abundance is progressively predicted at $\log_2 s$ levels where s is the scale factor. At the kth level, for the dilated feature reconstruction module in Fig. 3(b), we first extract the deep features by stacking convolutional layers simply. Given $\mathcal{A}_{k-1}$ as the input at the kth level, we have

$$\mathcal{F}_k = \mathcal{H}(\cdots(\mathcal{H}(\mathcal{A}_{k-1}))), \tag{11}$$

in which $\mathcal{F}_k$ represents deep features in LR space at this level, and $\mathcal{H}(\cdot)$ denotes the convolution operation followed by the ReLU activation function. In HSI, the abundance of material in a pixel is closely related to that of neighboring pixels (Irmak, Akar, & Yuksel, 2018). As the SR network processes the abundance coefficient instead of the pixel value, the large scale dependency should be taken into consideration. Therefore, to obtain a larger receptive field in limited layers, we exploit dilated convolutions to acquire more informative features. The dilated convolution is known for its expansion capacity of the receptive field while keeping the merit of the traditional convolution. Specifically, the 2D dilated convolution is implemented by inserting 'holes' (zeros) in the convolutional kernel without adding parameters. For a convolution kernel with size 3×3, the size of the resulting dilated filter is $(2d+1)\times(2d+1)$ with the dilation rate d. In standard convolution, $d = 1$. When $d > 1$, the dilated convolution would have a larger receptive field with the same parameters. However, using the same dilation rate for all layers may lead to gridding, losing a large portion of information. To avoid such deterioration, the hybrid dilated convolution (Wang et al., 2018) is adopted to cover a square region without any holes or missing edges. Practically, seven layers are stacked to extract deep features, which follows a sawtooth wave-like repeat. Hence, the equivalent receptive field of each layer is 3, 5, 7, 9, 7, 5, and 3, respectively. Consequently, it can be easily obtained that the receptive field of the layers is 33×33. If the traditional 3×3 convolution layer is used, the network will either have a receptive field of size 15×15 with the same network depth (*i.e.,* 7) or have a depth of 16 with the same receptive field (*i.e.,* 33×33). In this way, our model could make a trade-off between the size of the receptive field and network depth while fully exploiting the spatial information among abundance maps. For the sake of improving the spatial resolution, according to the upscale level, an upsampling module with scale 2 is used here to map the deep representations into HR space. Thus, the upscaled residual is given by

$$\mathcal{R}_k = \mathcal{U}(\mathcal{F}_k), \tag{12}$$

where $\mathcal{U}(\cdot)$ corresponds to the upsampling operation with the basic factor 2. We leverage the PixelShuffle (Shi et al., 2016) operator followed by a convolutional layer to conduct the upsampling procedure. The intermediate abundance map $\mathcal{A}_k$ is reconstructed in a residual way, which is

$$\mathcal{A}_k = \mathcal{A}_{k-1} \uparrow + \mathcal{R}_k, \tag{13}$$

where $\mathcal{A}_{k-1} \uparrow$ refers to the bicubic upscale version of the input. The output of this module can be regarded as the intermediate result with the upscale factor 2 and fed to the next level for further reconstruction. Different from Lai et al. (2017), our model does not constrain the intermediate result in a supervised manner. Hence, this progressive structure attempts to ease the learning process, especially at a large scale factor.

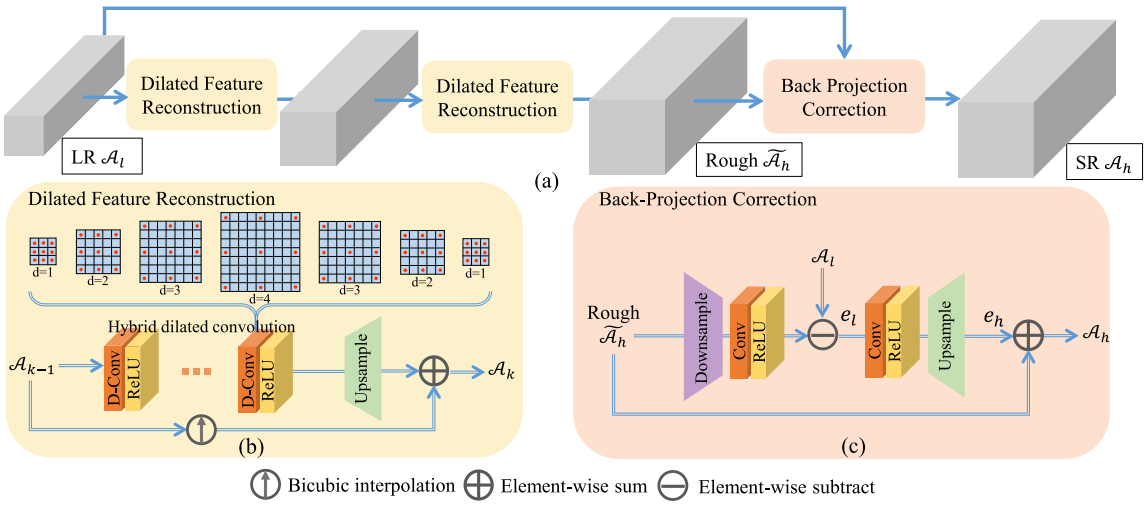

Fig. 3. The whole structure of the dilated projection correction network at the scale factor 4.

4.2.2. Back-projection correction

In most cases, the SR methods often deploy the upsampling module as the last part of the end-to-end networks. On the one hand, we transform the SR problem into the abundance domain but the generated abundance representation is not directly supervised. On the other hand, the decoder preserving the endmember spectrum reconstructs both LR and HR HSIs. Therefore, the HR abundance map should be enforced to have a similar spectral pattern as the LR one. In this case, we design the back-projection correction module to constrain the abundance representation implicitly, which is inspired by the classical method of iterative back-projection (Irani & Peleg, 1991). In this correction module, the difference between the downsampled rough abundance and the original LR one is back-projected into HR space to correct the distortion in a residual way. Meanwhile, we learn the linear combination of all fractions at each position to exploit the spectrum correlation among neighboring abundance values. Thus, the LR error would guide the distortion correction from the spectrum aspect. With the learnable correction strategy, the generated HR abundance representation could be adjusted to keep a similar pattern to the LR counterpart, which could reduce the distortion caused by the same spectrum reconstruction.

The detailed structure of the back-projection correction module is displayed in Fig. 3(c). Taking the rough HR representation $\tilde{\mathcal{A}}_h$ as input, this module first projects the super-resolved abundance into LR space, that is

$$\mathcal{S} = \mathcal{H}(\mathcal{D}(\tilde{\mathcal{A}}_h)), \tag{14}$$

where $\mathcal{S}$ refers to the downsampled counterpart of the rough abundance in HR space, $\mathcal{D}$ corresponds to the downsampling operation and $\mathcal{H}$ refers to one convolutional layer with ReLU. In this paper, we employ the reverse operation of PixelShuffle followed by a convolutional layer to conduct the downsampling procedure. To correct the distortion, the difference between the observed LR map $\mathcal{S}$ and the original $\mathcal{A}_l$ is obtained by the element-wise subtract as follows:

$$e_l = \mathcal{A}_l - \mathcal{S}. \tag{15}$$

Ideally, the difference between the projected LR map $\mathcal{S}$ and the original $\mathcal{A}_l$ is expected to be as small as possible. In this case, the downsamlped version of the rough abundance could be equal to the LR abundance. Then, we map the difference to the HR space to acquire the residual for back correction:

$$e_h = \mathcal{U}(\mathcal{H}(e_l)), \tag{16}$$

where $\mathcal{U}$ corresponds to the upsampling operation and $\mathcal{H}$ refers to one convolution layer with ReLU. Finally, the rough prediction could be corrected by

$$\mathcal{A}_h = \tilde{\mathcal{A}}_h + e_h. \tag{17}$$

Through the back-projection correction strategy, the LR and HR embeddings could have similar patterns for predicting pixel values by the same endmember spectrum. Similarly, we apply the ReLU function and normalization on the generated abundance in HR space to satisfy the nonnegative and sum-to-one constraints.

4.3. Loss function

As the HSI data have both spatial and spectral characters, we employ a joint function to constrain the predictive image similar to the ground truth, which is defined as the reconstruction loss term:

$$\mathcal{L}_{rec} = \mathcal{L}_{spatial} + \lambda \mathcal{L}_{spectral}, \tag{18}$$

where λ is a parameter that balances the trade-off between spatial loss and spectral difference. Since the l_1 loss has been proved to be more effective than the l_2 loss (Lim et al., 2017; Zhao, Gallo, Frosio, & Kautz, 2016), the spatial loss is constructed by l_1 norm, which is defined as follow:

$$\mathcal{L}_{spatial} = \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{I}^n - \hat{\mathcal{I}}^n\|_1, \tag{19}$$

where $\hat{\mathcal{I}}^n$ and $\mathcal{I}^n$ are the reconstructed HSI and ground-truth HSI, respectively, and N denotes the number of images in one training batch.

We utilize a measurement (Li et al., 2018) based on the spectral angle mapper (SAM) (Kruse et al., 1993) criterion to assess the spectral similarity, that is:

$$\mathcal{L}_{spectral} = \frac{1}{\pi}\frac{1}{N}\sum_{n=1}^{N} \arccos\left(\frac{\mathcal{I}^n \cdot \hat{\mathcal{I}}^n}{\|\mathcal{I}^n\|_2 \|\hat{\mathcal{I}}^n\|_2}\right), \tag{20}$$

where $\cdot$ is the tensor product and $\|\cdot\|_2$ is the tensor norm. By constraining both two aspects, we can considerably reduce both the numerical error and spectral distortion in the meanwhile.

Algorithm 1: Training procedure of aeDPCN

Input: Low-resolution HSI X
Output: Super-resolution result HSI $\hat{Y}$
For k_1 epochs **do**
  Evaluate loss Eqs. (21);
  Update parameters of auto-encoder by AdamOptimizer: $\nabla(\mathcal{L}_1)$;
end
Freeze the parameters of auto-encoder and use it to guide the training of DPCN
For k_2 epochs **do**
  Evaluate loss Eqs. (22);
  Update parameters of DPCN by AdamOptimizer: $\nabla(\mathcal{L}_2)$;
end
For k_3 epochs **do**
  Evaluate loss Eqs. (23);
  Update parameters of aeDPCN by AdamOptimizer: $\nabla(\mathcal{L}_3)$;
end
Return $\hat{Y} = \mathcal{M}_{aeDPCN}(X)$;

4.4. Training procedure

Since the proposed architecture is made up of two parts with different functions, we optimize the network with back-propagation following the procedure described below. The whole procedure is summarized in Algorithm 1.

Step 1: We first update the autoencoder network to decompose and reconstruct the HSI data. Accordingly, the training target of this step is to minimize the following objective:

$$\mathcal{L}_1 = \mathcal{L}_{rec}(\hat{\mathcal{I}}, \mathcal{I}), \tag{21}$$

in which $\hat{\mathcal{I}}$ and $\mathcal{I}$ are the ground truth HSI and reconstructed HSI respectively. In this step, the autoencoder network is ideally expected to reconstruct the original input as completely as possible. Since the autoencoder network is independent of the SR network, we use the training data at all scales in both LR and HR space to train this autoencoder just once. With this data augmentation strategy, the autoencoder is equipped with scale invariance.

Step 2: The estimated weights of the autoencoder network are fixed. Sharing the same decoder, we update the parameters of the SR network to generate both abundance representation and HR prediction. Thus, in this step, we expect that the generated abundance representation not only preserves the spectral information but also has a high spatial resolution to reconstruct satisfied HR HSI. Accordingly, the loss function of this step is:

$$\mathcal{L}_2 = \mathcal{L}_{rec}(\hat{\mathcal{Y}}, \mathcal{Y}). \tag{22}$$

Step 3: We combine the objective functions of the above two steps to jointly fine-tune the whole network so that the two networks can promote each other to produce better results. Consequently, the objective function of this step is composed by:

$$\mathcal{L}_3 = \mathcal{L}_{rec}(\hat{\mathcal{X}}, \mathcal{X}) + \mathcal{L}_{rec}(\hat{\mathcal{Y}}, \mathcal{Y}). \tag{23}$$

4.5. Implementation details

According to the datasets, both the input and output nodes of the proposed network are the same number of spectral bands in HSI data, denoted as L. We set the filter size of all convolutional layers to 3×3 except for that in the back-projection correction module, where the kernel size is set to 1×1. To ensure that the size of the feature map is not changed, the zero-padding strategy is applied for these convolutional layers with kernel size 3×3. For the dilated reconstruction module, the convolutional layers for feature extraction have 64 filters, except for the last layer, *i.e.*, the output channel of this layer is identical to the number of endmember spectrum c for residual learning. For the loss function, we set $\lambda = 0.1$. For the training phase, we empirically choose a mini-batch size of 16, and use Adam optimizer with weight decay of $1e{-}4$. The initial learning rate is set as 0.0001 and is decayed by 10 times after 2000 epochs while the total epoch is 3000. Our model is implemented by PyTorch on NVIDIA GTX 1080Ti.

4.6. Image quality metrics

For quantitative comparison, we use four quality measures to evaluate the prediction HR HSI $\hat{\mathcal{Y}}$ with its corresponding ground truth image $\mathcal{Y}$, including peak signal-to-noise ratio (PSNR), erreur relative globale adimensionnelle de synthèse (ERGAS) (Zeng, Huang, Liu, Zhang, & Zou, 2010), spectral angle mapper (SAM) (Kruse et al., 1993), and structural similarity (SSIM) (Wang, Bovik, Sheikh, Simoncelli, et al., 2004).

The widely used PSNR index is the mean ratio between the maximum power of the image and the power of the residual errors of all the spectral bands. For the ith spectral band, the PSNR is calculated by:

$$\mathrm{PSNR}(\mathcal{Y}_i, \hat{\mathcal{Y}}_i) = 10 \cdot \log_{10}\left(\frac{\max(\mathcal{Y}_i)^2}{mean\|\mathcal{Y}_i - \hat{\mathcal{Y}}_i\|_2^2}\right). \tag{24}$$

A higher PSNR value indicates a better image quality of the reconstructed HSI.

ERGAS measures the band-wise normalized root of mean square error (RMSE) between the reference HSI $\mathcal{Y}$ and the reconstructed HSI $\hat{\mathcal{Y}}$, with the best value at 0. It is defined as:

$$\mathrm{ERGAS}(\mathcal{Y}, \hat{\mathcal{Y}}) = \frac{100}{s}\sqrt{\frac{1}{L}\sum_{i=1}^{L}\frac{mean\|\mathcal{Y}_i - \hat{\mathcal{Y}}_i\|_2^2}{mean(\mathcal{Y}_i)^2}}, \tag{25}$$

where s is the scale factor between the LR HSI and HR HSI, and L is the number of spectral bands in HSI data.

SAM is commonly used to quantify the spectral information preservation at each pixel. More precisely, this index calculates the angle between two vectors of the estimated and reference spectra to represent spectral similarity. The SAM values near zero indicate high spectral similarity with no spectral distortion. The SAM is defined as:

$$\mathrm{SAM}(\mathcal{Y}, \hat{\mathcal{Y}}) = \arccos\left(\frac{\mathcal{Y} \cdot \hat{\mathcal{Y}}}{\|\mathcal{Y}\|_2\|\hat{\mathcal{Y}}\|_2}\right). \tag{26}$$

Another well-known index is the SSIM. For the ith spectral band, it is defined as:

$$\mathrm{SSIM}(\mathcal{Y}_i, \hat{\mathcal{Y}}_i) = \frac{(2\mu_{\mathcal{Y}_i}\mu_{\hat{\mathcal{Y}}_i} + c_1)(2\sigma_{\mathcal{Y}_i,\hat{\mathcal{Y}}_i} + c_2)}{(\mu_{\mathcal{Y}_i}^2 + \mu_{\hat{\mathcal{Y}}_i}^2 + c_1)(\sigma_{\mathcal{Y}_i}^2 + \sigma_{\hat{\mathcal{Y}}_i}^2 + c_2)}, \tag{27}$$

where $\mu_{\mathcal{Y}_i}$ and $\mu_{\hat{\mathcal{Y}}_i}$ are the means of $\mathcal{Y}_i$ and $\hat{\mathcal{Y}}_i$, respectively, $\sigma_{\mathcal{Y}_i}$ and $\sigma_{\hat{\mathcal{Y}}_i}$ are the variances of $\mathcal{Y}_i$ and $\hat{\mathcal{Y}}_i$, respectively, $\sigma_{\mathcal{Y}_i,\hat{\mathcal{Y}}_i}$ is the covariance of $\mathcal{Y}_i$ and $\hat{\mathcal{Y}}_i$, c_1 and c_2 are constants set as 0.0001 and 0.0009, respectively. The best value of SSIM is 1. The mean SSIM is estimated by averaging the SSIM values of all bands in HSI data.

5. Experiments and results

5.1. Datesets and experimental setup

The proposed method is evaluated on the two benchmark datasets: Chikusei (Yokoya & Iwasaki, 2016) and Pavia datasets.

Table 2
The trade-off between performance and parameter on the number of endmember spectra in the proposed aeDPCN evaluated on the testing set of Pavia dataset at the scale factor 4.

Number	PSNR	SAM	Params $\times 10^6$	FLOPs $\times 10^9$
10	27.75	6.25	0.07	0.71
20	28.01	5.90	0.26	2.73
30	28.44	5.44	0.58	5.99
40	28.50	5.40	1.03	10.96

Table 3
Ablation study. Quantitative comparisons among some variants of the proposed aeDPCN method over the testing set of Chikusei dataset at the scale factor 4.

Models	PSNR↑	ERGAS↓	SAM↓	SSIM↑
Our-w/o DC	38.86	5.36	2.96	0.90
Our-w/o BPC	39.04	5.47	3.16	0.91
Our-w/o FT	38.92	5.44	3.02	0.90
aeDPCN	39.12	5.22	2.89	0.92

The Chikusei dataset is an airborne hyperspectral dataset taken by the Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Ibaraki, Japan. The hyperspectral scene consists of 2517×2335 pixels with 128 bands in the spectral range from 363 nm to 1,018 nm. We first extract the top region of this image to form the testing data, which has four non-overlap hyperspectral images with $400 \times 400 \times 128$ pixels. The remaining region of this image are clipped into 400×400 sub-images with the overlap of 200 pixels. Ten percent of total sub-images are randomly divided as validation datasets. In order to enrich the training dataset, the rest sub-images are augmented by randomly rotating and flipping with a probability of 0.5.

The Pavia dataset is a real hyperspectral scene acquired by the ROSIS sensor during a flight campaign over Pavia. Since we discard the last band of the Pavia University for unity, both Pavia Centre and Pavia University are employed to extend the dataset with 102 spectral bands. The size of Pavia Centre image is $1,096 \times 1,096$ pixels with missing information in the center part of the image so we discard the missing part and use the left part which has $1,096 \times 220$ valid pixels for validation and test without overlapping. The rest part of subimages is extracted for training in the experiments. The Pavia University scene contains 610×340 pixels and we select the sub-region with the richest texture details as the test image, 220×220 pixels as shown at the first row of Fig. 4. All sub-images for training are cropped into 220×220 with the overlap of 110 pixels. To extend the training dataset, the training data is also augmented by the same way.

Both scenes are regarded as ground-truth of HR HSI and normalized into between 0 and 1. In order to simulate LR HSI, these HR images are downsampled by bicubic interpolation to obtain the corresponding counterparts. There is no additive noise during the degradation process.

Six strategies are chosen as the baseline for comparison: bicubic, exSRCNN, GDRRN (Li et al., 2018), FCNN (Mei et al., 2017), SSIN (Zheng et al., 2019) and MCNet (Li et al., 2020). The exSRCNN denotes the extended SRCNN since the original SRCNN (Dong et al., 2015) is designed for gray image SR, we adjust the first and last layers adapted to the training data. With the available source codes in public, the FCNN, GDRRN and MCNet algorithms are trained by their best settings. We reimplement the SSIN method according to the original paper as best as possible. For comparison, we use the average results of four quantitative indexes to evaluate the proposed method at the upscaling factors 2, 4 and 8.

5.2. Ablation studies

Number of endmember spectra. The most important parameter of the whole architecture is the number of endmember spectra c. As basis spectra carry the spectral information of the material, this number determines the representation capacity of the proposed structure and the computational complexity. We train the proposed model on the Pavia dataset with different settings, *e.g.*, $c = 10, 20, 30, 40$ at the upscale factor 4, and show the trade-off between performance and parameter in Table 2. According to the results, increasing the spectrum number, the performance of the network has improved gradually with the increased computational complexity. When we increase the number of endmember spectra from 30 to 40, there is only a little improvement in both spatial and spectral aspects. Therefore, we choose $c = 30$ for Pavia dataset to strike a balance between performance and computational complexity.

Dilated convolution. In order to acquire more information in limited layers, we deploy dilated convolutions for feature extraction. To demonstrate the effect of the dilated convolution, we replace dilated convolutions by standard convolutions with the kernel size of 3×3 and obtain the variant of our method, *i.e.,* Our-w/o DC. When the depth in the dilated feature reconstruction module is 7, the variant has a receptive field of size 15×15. As reported in Table 3, our method with dilated convolutions achieves better performance on all indices especially for the PSNR value (+0.26 dB). Owing to the dilated convolution, our method could acquire deep spatial information in limited layers, which contributes to a light-weight model with comparable performance.

Back-projection correction. In view of sharing the same decoder that preserves the spectral information, the super-resolved abundance should have a similar pattern as the LR one. Thereby, we design the back-projection correction module to correct the spectral distortion. To verify the effectiveness of the back-projection correction module, we compare the performance of the method with and without this module. As shown in Table 3, with the back-project correction mechanism, our method has achieved a significant performance gain on the spectral aspect compared to Our-w/o BPC. Since we use the 1×1 convolution in this block, the back-projection strategy focuses on the spectral dimension. Although it has a relatively small improvement on the index of spatial reconstruction error (*i.e.*, PSNR), this correction module could significantly reduce the spectral distortion.

Fine-tuning. Since the proposed method consists of two parts: autoencoder-based spectral decomposition, dilated projection correction network for super-resolution, which are unified in an end-to-end manner, we optimize the whole network by fine-tuning the pretrained two parts. To some extent, this strategy reduces the difficulty of network training. The two parts could promote each other to achieve better performance. We train the proposed method from scratch with the objective of step 3 and tabulate the results in Table 3, corresponding to Our-w/o FT. It can be seen that without fine-tuning, the performance of the proposed method suffers from deterioration for both spatial and spectral evaluation.

5.3. Qualitative results

To demonstrate the effectiveness of our proposed SR method for HSI, we first exhibit the reconstruction results of different compared methods with ground truth image in the last column in Figs. 4 and 5. As for the Chikusei dataset, we choose two sub-images from the test set, representing agricultural and urban areas and the 84th band of them are visualized in Fig. 5. In Fig. 4, the first row shows the 82th bands of a sub-region from Pavia University scene and we select 48th band of a sub-region from Pavia Centre scene in the second row of Fig. 4. For

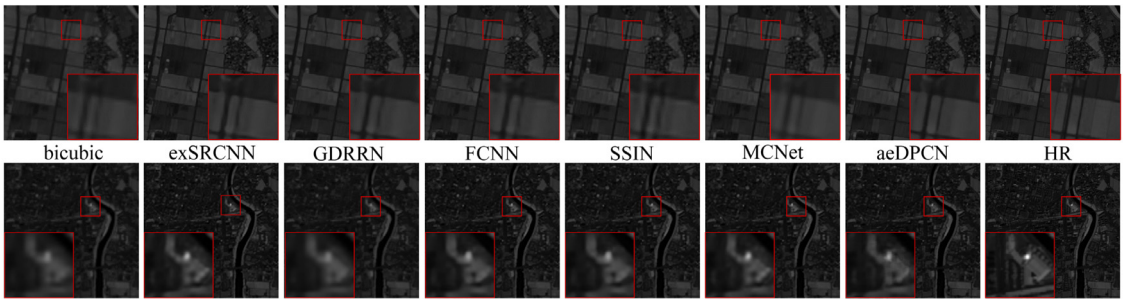

Fig. 4. Visual comparison for ×4 HSI SR on two representative sub-images from Chikusei dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

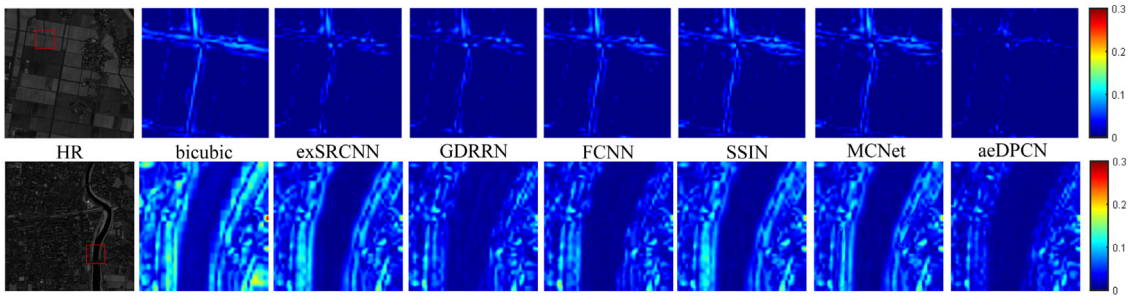

Fig. 5. Spatial absolute difference for ×4 HSI SR on two representative sub-images from Chikusei dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

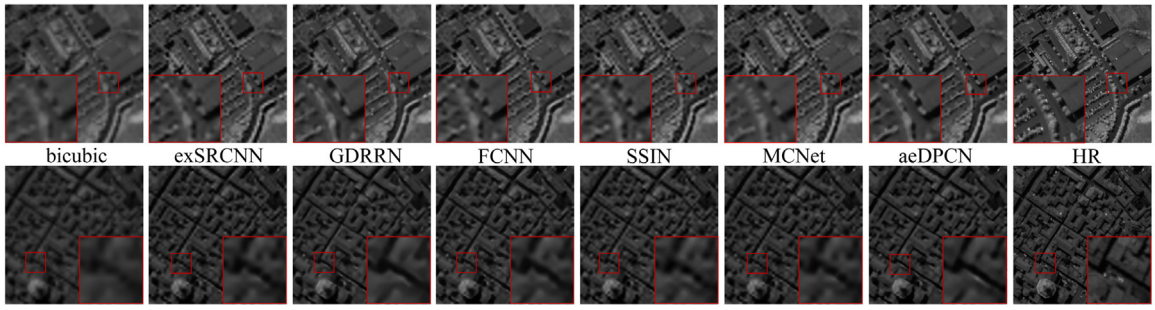

Fig. 6. Visual comparison for ×4 HSI SR on the test images from Pavia University and Pavia Centre.

better visualization, we highlight a small area in the red box. We can observe that our method could recover sharper and clearer edges or outlines which are blurry in other reconstruction results. Specifically, our method behaves well in restoring the building boundaries in the scene. Due to the fact that our method predicts the individual pixel value from the abundance domain by multiplying the spectral representations, they might be more accurate in composing sharp scene information. For the sake of comparing the reconstruction results over all bands, in Figs. 6 and 7, we display the mean absolute difference over all bands between reconstructed result and the ground truth for two datasets. Since the original scenes are not very clear, we only display a small area of each scene. As illustrated in Figs. 6 and 7, our results recover HR images with less reconstruction error.

5.4. Quantitative results

Tables 4 and 5 demonstrate the experimental results of six methods evaluated on the Chikusei and Pavia test sets, including PSNR, ERGAS, SAM, and SSIM. It can be noticed that the exSRCNN method suffers from spectral distortion in both datasets, more severe than the traditional interpolation method. Since the exSRCNN is not designed for the HSI SR, it neglects the spectral information, which is the key characteristic in HSI data. Even if the PSNR value of exSRCNN is higher than that of the bicubic, exSRCNN does not achieve competitive performance on the spectral information. As the SSIN method reconstructs the HR data in a coarse-to-fine manner, it can acquire good PSNR scores. However, as indicated in the table, when the upsampling scale is large (*e.g.*, 8), the performance of SSIN evaluated by SAM is not promising. Although the mixed convolutions in MCNet are designed to mine the spatial information, the spectral information is not well preserved.

Based on the experimental results, equipped with the dilated convolution and back-projection correction, the proposed aeDPCN outperforms the others in terms of all quantitative indexes, and it is quite stable at different upscaling factors for both datasets. Specifically, since the original Pavia scenes do not have high quality, the improvement on test set is relatively small. On the contrary, this demonstrates that our method can handle challenging scenarios much better than other methods. In addition, at the upscaling factor 8, our method has much improved performance on the SAM value, which shows effectiveness in preserving the spectral information with the improvement of spatial resolution. This is mainly because we enhance the spatial resolution in the abundance domain with dilated convolutions

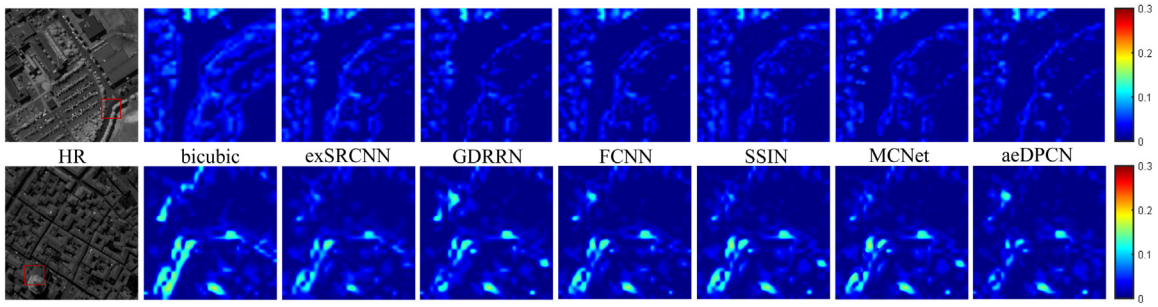

Fig. 7. Spatial absolute difference for ×4 HSI SR on the test images from Pavia University and Pavia Centre.

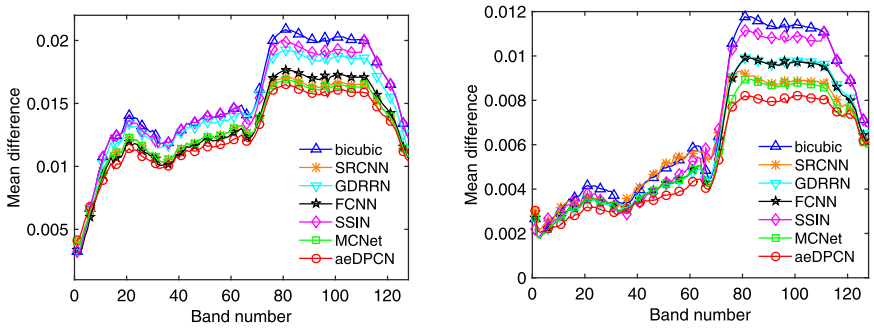

Fig. 8. Absolute difference along spectral dimension of the ×4 HSI SR on two representative sub-images from Chikusei testset.

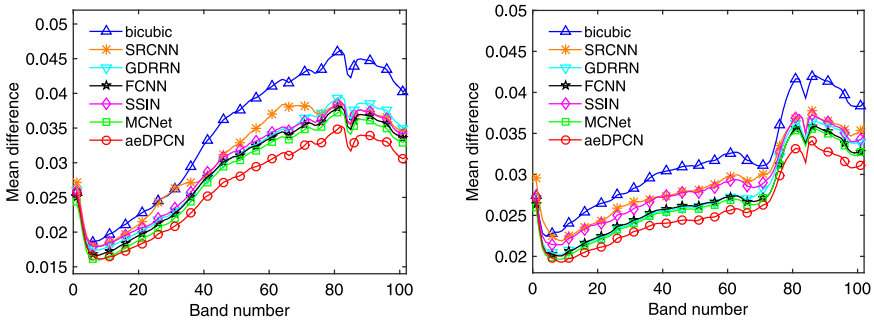

Fig. 9. Absolute difference along spectral dimension of the ×4 HSI SR on the test images from Pavia University and Pavia Centre.

Table 4
Quantitative evaluation on Chikusei dataset of state-of-the-art HSI SR algorithms: average PSNR/ERGAS/SAM/SSIM for scale factors 2, 4 and 8. bold indicates the best result.

Scale	Index	bicubic	exSRCNN	GDRRN	FCNN	SSIN	MCNet	aeDPCN
×2	PSNR↑	42.35	43.62	43.01	45.08	44.46	45.22	**45.51**
	ERGAS↓	7.64	6.43	7.02	5.94	6.70	5.88	**5.64**
	SAM↓	2.37	2.44	2.07	2.19	2.32	2.20	**1.87**
	SSIM↑	0.96	0.97	0.97	0.97	0.97	**0.98**	**0.98**
×4	PSNR↑	37.35	37.51	38.07	38.65	37.92	38.59	**39.12**
	ERGAS↓	6.59	6.85	6.10	5.69	6.42	5.94	**5.22**
	SAM↓	3.93	4.05	3.54	3.62	3.86	3.35	**2.89**
	SSIM↑	0.88	0.90	0.90	0.91	0.90	0.90	**0.92**
×8	PSNR↑	34.45	34.55	34.42	35.06	34.62	34.88	**35.49**
	ERGAS↓	4.54	4.54	4.57	4.25	4.32	4.37	**4.04**
	SAM↓	5.61	5.79	5.34	5.23	5.82	5.46	**4.98**
	SSIM↑	0.82	0.83	0.82	0.83	0.82	0.82	**0.84**

Table 5
Quantitative evaluation on Pavia dataset of state-of-the-art HSI SR algorithms: average PSNR/ERGAS/SAM/SSIM for scale factors 2, 4 and 8. bold indicates the best result.

Scale	Index	bicubic	exSRCNN	GDRRN	FCNN	SSIN	MCNet	aeDPCN
×2	PSNR↑	30.78	31.27	32.47	33.61	33.51	33.84	**34.02**
	ERGAS↓	9.34	9.24	8.12	6.83	6.92	6.75	**6.56**
	SAM↓	4.58	7.05	4.23	3.92	4.31	4.12	**3.53**
	SSIM↑	0.87	0.91	0.91	0.93	0.93	0.93	**0.95**
×4	PSNR↑	26.71	27.43	27.74	27.89	27.74	27.96	**28.44**
	ERGAS↓	7.35	6.82	6.55	6.44	6.56	6.46	**6.08**
	SAM↓	6.37	6.86	5.84	5.85	5.90	5.62	**5.44**
	SSIM↑	0.66	0.73	0.73	0.74	0.73	0.75	**0.78**
×8	PSNR↑	24.11	24.32	24.40	24.62	24.61	24.70	**24.90**
	ERGAS↓	4.93	4.82	4.71	4.65	4.65	4.56	**4.34**
	SAM↓	8.14	8.84	7.92	7.77	8.37	7.83	**7.32**
	SSIM↑	0.46	0.50	0.50	0.51	0.50	0.52	**0.53**

and the predictive abundance representations are back-projected to the LR space to correct the spectral difference. In the meanwhile, the spectrum information is preserved in the decoder as prior information, which is also remained constant at any factor for both LR and HR. In this way, our performance is more stable than all algorithms even at a large scale factor.

To further show our advantages in preserving the spectral information, we display the difference of the compared methods along the spectral dimension in Figs. 8 and 9 for the Chikusei

Table 6
Parameter and complexity comparison. The results are evaluated at the scale factor 4.

Methods	exSRCNN	GDRRN	FCNN	SSIN	MCNet	aeDPCN
Params $\times 10^6$	0.66	0.52	0.03	72.08	2.17	0.58
FLOPs $\times 10^9$	24.39	106.50	148.16	166.07	2075.71	5.99

and Pavia datasets, respectively, which is collected by averaging the whole bands of the selected images. Instead of displaying the spectral reflectance at several positions, these diagrams would be more representative. In Fig. 8, we show the absolute difference of two representative sub-images from Chikusei. For the Pavia dataset, the left one of Fig. 9 is the mean absolute difference of a test image from Pavia University scene while the right one is from Pavia Centre. In these figures, in view of the spectral aspect, our method, aeDPCN (red curve), has the lowest error for all bands. The basic idea of the proposed method is to extract spectral information by autoencoder and improve the spatial resolution with dilated convolutions and back-projection strategy in the abundance domain. Therefore, on the one hand, through the constant decoder carrying the spectral representation as prior information, it can be ensured that spectral properties are largely preserved during the SR process. On the other hand, owing to the back-projection correction, our method can avoid spectral distortion well for reconstruction.

5.5. Parameter and complexity

We use the original codes of compared methods to calculate the parameter and computational complexity. Table 6 concludes the results that are evaluated on the input size of 48×48 at scale 4. We can see that at the same level of parameters, our method has the smallest computational overhead, referring to the FLOPs. Even if FCNN has fewer parameters, the 3D convolution is more computationally intensive, which is the same case in MCNet. Due to the autoencoder-based the spectral decomposition, we transform the SR problem into the abundance domain by several fully-connected layers. Meanwhile, the dimension of the abundance representation is much smaller than that of original HSI which other methods have taken as input. Consequently, our method could significantly reduce the computational complexity.

5.6. Real image super-resolution

We also conduct SR experiments on real-world hyperspectral images. The performance is evaluated on Washington DC Mall data,[1] which is a 191-band airborne multispectral scanner dataset with a size of $1,208 \times 307$. We first crop the center region of the image to obtain a subimage with $1,200 \times 300 \times 191$ pixels, which are further divided into training and test data. Specifically, the left bottom region ($200 \times 200 \times 191$) is extracted as the testing data and we extract overlapping patches from the remaining region of the subimage as reference HR images for training. In this case, the original HR images are not available and the degradation model is unknown either. We compare our proposed aeDPCN with other state-of-the-art methods at the scale factor 4. As shown in Fig. 10, our aeDPCN recovers sharper edges and finer details than other state-of-the-art methods. These results further indicate the benefits of preserving endmember spectra as prior information, which is consistent for different or unknown degradation models.

[1] https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html.

6. Discussion

Generalization performance. Since we extract the endmember spectrum as the prior information for the SR problem, the test data of our method are required to have the same endmember spectrum as the training data. On the one hand, our proposed method can be regarded as an unsupervised image-specific SR method for a single HSI. As proposed in Shocher, Cohen, and Irani (2018), due to modest amounts of computational resources, our method does not require pretraining. Specifically, at the test time, the proposed image-specific aeDPCN is trained on examples extracted internally, from the test image itself. Therefore, our network can be adapted to the arbitrary degradations of HSI at test time. On the other hand, fortunately, most remotely sensed hyperspectral images are only one large HSI with internal similarities, satisfying this requirement.

Super-resolution vs. unmixing. Different from the previous works (Miao & Qi, 2007; Palsson et al., 2018; Qu & Qi, 2018; Su et al., 2019) involving unmixing based on matrix factorization, our method unifies the spectral decomposition and SR process into an end-to-end model to realize the single-image SR for HSI, instead of unmixing. Since we transform the SR problem into the abundance domain, the hyperparameter c not only has a specific physical meaning, *i.e.,* the number of endmembers, but also determines the expressive power and complexity of the network. As indicated in Table 2, if high-quality super-resolution is the goal, then it is generally better to set c larger than the number of actual materials in the scene. Meanwhile, in Miao and Qi (2007), it has been pointed out that the unsupervised NMF algorithms are sensitive to initializations due to the existence of local optima and the algorithm performance highly depends on the distance between the initial point and the global optimum. Hence, these works (Palsson et al., 2018; Qu & Qi, 2018; Su et al., 2019) for HSI unmixing are dedicated to imposing the special constraints and the optimal initialization to achieve good performance on unmixing. However, except for simplex constraints (Eqs. (6) and (7)) for stability (Lanaras et al., 2015), we impose neither strong constraints nor initializations on the abundance or the endmember. Therefore, the endmember extracted in our method might be not correct in the physical sense but more like a primitive (or an atom/base of a dictionary) with strong expression ability.

Future work. The proposed deep-learning-based method may largely rely on supervised learning. If there are only a few training pairs but high spectral dimensionality of hyperspectral image data, it would be easy to cause the over-fitting problem. Therefore, in future research, we plan to combine the unsupervised learning (Ieracitano et al., 2020; Zhu, Fu, & Zhao, 2020) to provide more trainable samples. In addition, the meta-learning strategy based on our abundance SR framework would also be studied for better SR performance.

7. Conclusion

In this paper, we have proposed a novel dilated projection correction network with an autoencoder for HSI SR, called aeDPCN. In our method, we transform the SR problem into the abundance domain based on the insight that the HR and LR HSIs share the same basic spectrum, which could considerably reduce the computational complexity. The spatial resolution of the LR abundance map is then progressively improved by the dilated projection correction network with a large reception field. Meanwhile, we use the back-projection correction strategy to ensure the generated abundance to follow the same pattern as LR abundance, which can reduce the spectral distortion. The final prediction is produced by the same decoder which preserves the spectrum information. The qualitative and quantitative results on real HSI scenes have demonstrated the stability and superiority of our method over the state-of-the-art at different scale factors.

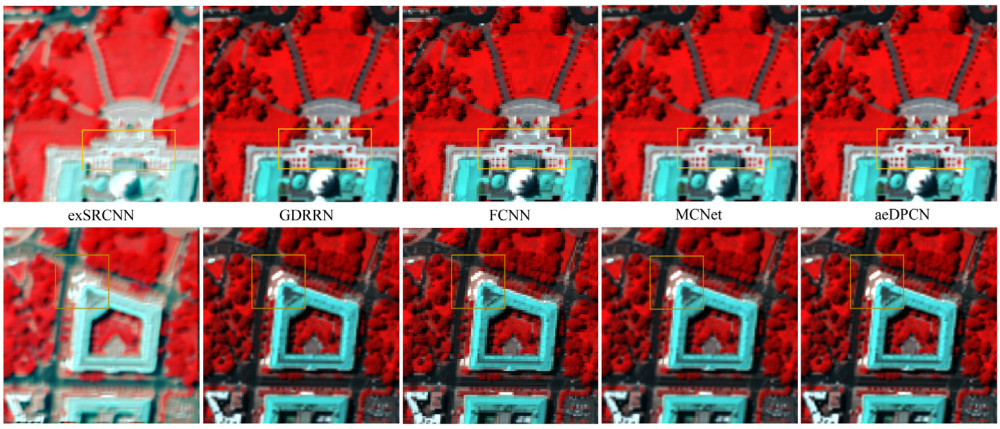

Fig. 10. Visual comparison for ×4 HSI SR on the real-world images from Washington DC Mall with spectral bands 60-27-17 as R-G-B.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the National Natural Science Foundation of China under Grant nos. 61773295 and 61971165, and the Natural Sciences and Engineering Research Council of Canada under Grant no. RGPIN-2020-04661.

References

Akhtar, N., Shafait, F., & Mian, A. (2015). Bayesian sparse representation for hyperspectral image super resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3631–3640).

Dong, W., Fu, F., Shi, G., Cao, X., Wu, J., Li, G., et al. (2016). Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Transactions on Image Processing*, *25*(5), 2337–2352.

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307.

Du, B., & Zhang, L. (2014). A discriminative metric learning based anomaly detection method. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(11), 6844–6857.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Haris, M., Shakhnarovich, G., & Ukita, N. (2018). Deep back-projection networks for super-resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1664–1673).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

He, S., Zhou, H., Wang, Y., Cao, W., & Han, Z. (2016). Super-resolution reconstruction of hyperspectral images via low rank tensor modeling and total variation regularization. In *Proceedings of the ieee international geoscience and remote sensing symposium* (pp. 6962–6965).

Hu, J., Li, Y., & Xie, W. (2017). Hyperspectral image super-resolution by spectral difference learning and spatial error correction. *IEEE Geoscience and Remote Sensing Letters*, *14*(10), 1825–1829.

Hu, J., Zhao, M., & Li, Y. (2019). Hyperspectral image super-resolution by deep spatial-spectral exploitation. *Remote Sensing*, *11*(10), 1229.

Huang, H., Yu, J., & Sun, W. (2014). Super-resolution mapping via multi-dictionary based sparse representation. In *Proceedings of the ieee international conference on acoustics, speech and signal processing* (pp. 3523–3527).

Ieracitano, C., Paviglianiti, A., Campolo, M., Hussain, A., Pasero, E., & Morabito, F. C. (2020). A novel automatic classification system based on hybrid unsupervised and supervised machine learning for electrospun nanofibers. *IEEE/CAA Journal of Automatica Sinica*, *8*(1), 64–76.

Irani, M., & Peleg, S. (1991). Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, *53*(3), 231–239.

Irmak, H., Akar, G. B., & Yuksel, S. E. (2018). A MAP-based approach for hyperspectral imagery super-resolution. *IEEE Transactions on Image Processing*, *27*(6), 2942–2951.

Jiang, J., Sun, H., Liu, X., & Ma, J. (2020). Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging*, *6*, 1082–1096.

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016a). Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1646–1654).

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016b). Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1637–1645).

Kruse, F. A., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., et al. (1993). The spectral image processing system (SIPS) interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, *44*(2–3), 145–163.

Lai, W.-S., Huang, J.-B., Ahuja, N., & Yang, M.-H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 624–632).

Lanaras, C., Baltsavias, E., & Schindler, K. (2015). Hyperspectral super-resolution by coupled spectral unmixing. In *Proceedings of the ieee international conference on computer vision* (pp. 3586–3594).

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4681–4690).

Li, Y., Hu, J., Zhao, X., Xie, W., & Li, J. (2017). Hyperspectral image super-resolution using deep convolutional neural network. *Neurocomputing*, *266*, 29–41.

Li, Q., Wang, Q., & Li, X. (2020). Mixed 2d/3d convolutional network for hyperspectral image super-resolution. *Remote Sensing*, *12*(10), 1660.

Li, Y., Zhang, L., Dingl, C., Wei, W., & Zhang, Y. (2018). Single hyperspectral image super-resolution with grouped deep recursive residual network. In *Proceedings of the ieee international conference on multimedia big data* (pp. 1–4).

Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 136–144).

Liu, T., Gu, Y., Chanussot, J., & Dalla Mura, M. (2017). Multimorphological superpixel model for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(12), 6950–6963.

Liu, D., Wen, B., Fan, Y., Loy, C. C., & Huang, T. S. (2018). Non-local recurrent network for image restoration. In *Advances in neural information processing systems* (pp. 1673–1682).

Mei, S., Yuan, X., Ji, J., Zhang, Y., Wan, S., & Du, Q. (2017). Hyperspectral image spatial super-resolution via 3D full convolutional neural network. *Remote Sensing*, *9*(11), 1139.

Miao, L., & Qi, H. (2007). Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(3), 765–777.

Palsson, B., Sigurdsson, J., Sveinsson, J. R., & Ulfarsson, M. O. (2018). Hyperspectral unmixing using a neural network autoencoder. *IEEE Access*, *6*, 25646–25656.

Plaza, A., Du, Q., Bioucas-Dias, J. M., Jia, X., & Kruse, F. A. (2011). Foreword to the special issue on spectral unmixing of remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(11), 4103–4110.

Qu, Y., & Qi, H. (2018). uDAS: An untied denoising autoencoder with sparsity for spectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(3), 1698–1712.

Qu, Y., Qi, H., & Kwan, C. (2018). Unsupervised sparse Dirichlet-net for hyperspectral image super-resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2511–2520).

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., et al. (2016). Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1874–1883).

Shocher, A., Cohen, N., & Irani, M. (2018). "zero-shot" super-resolution using deep internal learning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3118–3126).

Simões, M., Bioucas-Dias, J., Almeida, L. B., & Chanussot, J. (2014). A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(6), 3373–3388.

Su, Y., Li, J., Plaza, A., Marinoni, A., Gamba, P., & Chakravortty, S. (2019). Daen: Deep autoencoder networks for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(7), 4309–4321.

Tai, Y., Yang, J., & Liu, X. (2017). Image super-resolution via deep recursive residual network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3147–3155).

Tai, Y., Yang, J., Liu, X., & Xu, C. (2017). Memnet: A persistent memory network for image restoration. In *Proceedings of the ieee international conference on computer vision* (pp. 4539–4547).

Tong, T., Li, G., Liu, X., & Gao, Q. (2017). Image super-resolution using dense skip connections. In *Proceedings of the ieee international conference on computer vision* (pp. 4799–4807).

Veganzones, M. A., Simoes, M., Licciardi, G., Yokoya, N., Bioucas-Dias, J. M., & Chanussot, J. (2015). Hyperspectral super-resolution of locally low rank images from complementary multisource data. *IEEE Transactions on Image Processing*, *25*(1), 274–288.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–610.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., et al. (2018). Understanding convolution for semantic segmentation. In *Proceedings of the ieee winter conference on applications of computer vision* (pp. 1451–1460).

Wang, X., Ma, J., & Jiang, J. (2021). Hyperspectral image super-resolution via recurrent feedback embedding and spatial-spectral consistency regularization. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., et al. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the european conference on computer vision workshop* (pp. 1–16).

Wei, Q., Bioucas-Dias, J., Dobigeon, N., & Tourneret, J.-Y. (2015). Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3658–3668.

Wen, B., Kamilov, U. S., Liu, D., Mansour, H., & Boufounos, P. T. (2018). DeepCASD: An end-to-end approach for multi-spectral image super-resolution. In *Proceedings of the ieee international conference on acoustics, speech and signal processing* (pp. 6503–6507).

Wu, H., & Prasad, S. (2017). Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, *27*(3), 1259–1270.

Xie, W., Jia, X., Li, Y., & Lei, J. (2019). Hyperspectral image super-resolution using deep feature matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(8), 6055–6067.

Xie, W., Lei, J., Liu, B., Li, Y., & Jia, X. (2019). Spectral constraint adversarial autoencoders approach to feature representation in hyperspectral anomaly detection. *Neural Networks*, *119*, 222–234.

Xie, W., Li, Y., Hu, J., & Chen, D.-Y. (2018). Trainable spectral difference learning with spatial starting for hyperspectral image denoising. *Neural Networks*, *108*, 272–286.

Xie, W., Li, Y., Lei, J., Yang, J., Li, J., Jia, X., et al. (2020). Unsupervised spectral mapping and feature selection for hyperspectral anomaly detection. *Neural Networks*, *132*, 144–154.

Yokoya, N., Grohnfeldt, C., & Chanussot, J. (2017). Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, *5*(2), 29–56.

Yokoya, N., & Iwasaki, A. (2016). *Airborne hyperspectral data over Chikusei*: *Tech. rep.*, Japan: Space Application Laboratory, University of Tokyo.

Yokoya, N., Yairi, T., & Iwasaki, A. (2011). Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(2), 528–537.

Yuan, Y., Zheng, X., & Lu, X. (2017). Hyperspectral image superresolution by transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(5), 1963–1974.

Zeng, Y., Huang, W., Liu, M., Zhang, H., & Zou, B. (2010). Fusion of satellite images in urban area: Assessing the quality of resulting images. In *Proceedings of the international conference on geoinformatics* (pp. 1–4).

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In *Proceedings of the european conference on computer vision* (pp. 286–301).

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2472–2481).

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, *3*(1), 47–57.

Zheng, K., Gao, L., Ran, Q., Cui, X., Zhang, B., Liao, W., et al. (2019). Separable-spectral convolution and inception network for hyperspectral image super-resolution. *International Journal of Machine Learning and Cybernetics*, 1–15.

Zhu, G., Fu, J., & Zhao, F. (2020). Unsupervised/supervised hybrid deep learning framework for low dose phase contrast imaging. In *Journal of physics: conference series, Vol. 1624*. IOP Publishing, Article 052026.