Deep RGB-driven Learning Network for Unsupervised Hyperspectral Image Super-resolution*

Zhe Liu[1,2][0000−0002−3756−2678] and Xian-Hua Han[1,3][0000−0002−5003−3180]

[1] Graduate School of Sciences and Technology for Innovation, Yamaguchi University, 1677-1 Yoshida, Yamaguchi, 753-8511, Japan
[2] a501wbu@yamaguchi-u.ac.jp
[3] hanxhua@yamaguchi-u.ac.jp

Abstract. Hyperspectral (HS) images are used in many fields to improve the analysis and understanding performance of captured scenes, as they contain a wide range of spectral information. However, the spatial resolution of hyperspectral images is usually very low, which limits their wide applicability in real tasks. To address the problem of low spatial resolution, super-resolution (SR) methods for hyperspectral images (HSI) have attracted widespread interest, which aims to mathematically generate high spatial resolution hyperspectral (HR-HS) images by combining degraded observational data: low spatial resolution hyperspectral (LR-HS) images and high resolution multispectral or RGB (HR-MS/RGB) images. Recently, paradigms based on deep learning have been widely explored as an alternative to automatically learn the inherent priors for the latent HR-HS images. These learning-based approaches are usually implemented in a fully supervised manner and require large external datasets including degraded observational data: LR-HS/HR-RGB images and corresponding HR-HS data, which are difficult to collect, especially for HSI SR scenarios. Therefore, in this study, a new unsupervised HSI SR method is proposed that uses only the observed LR-HS and HR-RGB images without any other external samples. Specifically, we use an RGB-driven deep generative network to learn the desired HR-HS images using a encoding-decoding-based network architecture. Since the observed HR-RGB images have a more detailed spatial structure and may be more suitable for two-dimensional convolution operations, we employ the observed HR-RGB images as input to the network as a conditional guide and adopt the observed LR-HS/HR-RGB images to formulate the loss function that guides the network learning. Experimental results on two HS image datasets show that our proposed unsupervised approach provides superior results compared to the SoTA deep learning paradigms.

Keywords: Hyperspectral image · Super-resolution · Unsupervised learning.

* Graduate School of Sciences and Technology for Innovation, Yamaguchi University.

1 Introduction

With hyperspectral (HS) imaging, detailed spectral direction traces and rich spectral features in tens or even hundreds of bands can be obtained at every spatial location in a scene, which can significantly improve the performance of different HS processing systems. Existing HS image sensors typically collect HS data at a low spatial resolution, which severely limits their wide applicability in the real world. Therefore, generating high-resolution hyperspectral (HR-HS) images by combining the degraded observational data: low-resolution hyperspectral (LR-HS) and high-resolution multispectral/RGB (HR-MS/RGB) images, known as HS super-resolution (HSI SR) images, has attracted great attention in the field of computer vision [29, 32], medical diagnosis [19, 21, 33], mineral exploration [24, 30] and remote sensing [3, 20, 23]. According to the reconstruction principles, HSI SR is mainly divided into two categories: traditional mathematical model-based methods and deep learning-based methods. In the following, we will describe these two types of methods in detail.

1.1 Traditional Mathematical Model-based Methods

In the past decades, most HSI SR methods have focused on studying various manually computed a prior parameters to develop a mathematical model and use optimization techniques to solve the problem. Specifically, such methods have focused on developing a mathematical formulation to model the process of degrading HR-HS images into LR-HS images and HR-RGB images. Since the known variables of the observed LR-HS/HR-RGB images are much smaller than the underestimated variables of the HR-HS images, this task is extremely challenging and direct optimization of the formulated mathematical model would lead to a very unstable solution. Therefore, existing effort usually exploits various priors to regularize the mathematical model, i.e., to impose constraints on the solution space. Depending on the priors to be investigated, existing studies are generally classified into three different approaches: spectral unmixing-based [16], sparse representation-based [6] and tensor factorization-based methods [4]. In the spectral un-mixing-based method, Yokoya et al. [31] proposed a coupled non-negative matrix decomposition (CNMF) method, which alternately blends LR-HS images and HR-RGB images to estimate HR-HS images. Recently, Lanaras et al. [16] proposed a similar framework to jointly extract two observed images by decomposing the original optimization problem into two constrained least squares problems. A similar framework was proposed by Dong et al. [6], which employed the alternating direction method of multipliers (ADMM) to solve the spectral hash model for the robust reconstruction of the base image of HR-HS images.

In addition, sparse representation is widely used as an alternative mathematical model for HSI SR, where a spectral dictionary is first learned from the observed HR-LS image, and then the sparse coefficients of the HR-RGB image are calculated to reconstruct the HR-HS image. For example, Zhao et al. [34]

used the K-SVD method to learn the dictionary, and then adopted the sparse matrix decomposition to combine the LR-HS and HR-RGB images to reconstruct the HR-HS image. Inspired by the spectral similarity of neighboring pixels in latent HS images, Akhtar et al. [1] proposed to perform group sparsity and non-negativity representations, while Kawakami et al. [15] used a sparse regularizer to decompose the spectral dictionary. In addition, Han et al. [9] proposed a non-negative sparse coding algorithm that effectively exploits pixel sparsity and non-local spatial similarity in HR-HS images. Furthermore, the tensor factorization-based approach was shown to be feasible for the HSI SR problem. Motivated by the inherent low dimensionality of spectral signatures and the 3D structure of HR-HS images, He et al. [13] employed matrix factorization to decomposed the HR-HS image into two low-rank constraint matrices and showed impressive super-resolution results. Despite some improvements achieved using manually designed priors, super-resolution performance tends to be unstable and sensitive to the content varying in the under-studying investigated images as well as may lead to significant spectral distortions due to the insufficient representation of empirically designed priors.

1.2 Deep Learning-based Methods

Deep Supervised Learning-based Methods Due to the high success of DCNN in different vision tasks, DCNN-based approaches have been proposed for HSI SR tasks to automatically learn specific priors for the latent HR-HS images. Han et al. [10] firstly conducted a pioneering work for merging the HS/RGB image to estimating the latent HR-HS image using deep learning network, which contained three simple convolutional layers but demonstrated very impressive performance, and then employed more complex CNN architectures such as ResNet and DenseNet [8] for performance improvement. Dian et al. [5] proposed an optimization and learning integration strategy for the fusion task by first solving the Sylvester equation and then exploring a DCNN-based approach to improve the initialization results. Han et al. [12] further investigated a multi-level and multi-scale spatial and spectral fusion network to effectively fuse the observed LR-HS and HR-RGB images with large spatial structure differences. Xie et al. [28] studied the MS/HS fusion network using a low-resolution imaging model and spectral low-rank property of the HR-HS images, and solved the proposed MS/HS fusion network using an approximate gradient strategy. In addition, Zhu et al. [35] investigated a lightweight deep neural network, dubbed as the progressive zero centric residual network (PZRes-Net), to achieve efficiency and performance in solving the HS image reconstruction problem. Despite the significant improvement in reconstruction performance, all the above DCNN-based approaches require training with large external datasets, including the degraded LR-HS/HR-RGB images and corresponding HR-HS images, which are difficult to collect, especially for HSI SR tasks.

Deep Unsupervised Learning-based Methods As mentioned above, in practice, it is very difficult to collect enough training triples, especially the latent

HR-HS images for a good-generalization CNN model. Therefore, the quality and quantity of the collected training datasets (external datasets) usually become the bottleneck of DCNN-based approaches. To alleviate the heavy reliance on external datasets, several deep unsupervised learning methods have been investigated to take advantage of the powerful modeling capabilities of deep networks [17]. Qu et al. [22] attempted to solve the HSI super-resolution problem with unsupervised learning strategy and developed an encoder-decoder architecture that exploits the approximate structure of the low-rank spectral prior in the latent HR-HS images. Although this approach does not require external training samples to construct a CNN-based end-to-end model for the recovery of the HR-HS images, it requires careful designing alternative optimization procedure for two sub-problems, and tends to produce unstable super-resolution results. Liu et al. [18] proposed a deep unsupervised fusion learning (DUFL) method to generate the HR-HS image from a random noisy input using the observed HR-RGB and LR-HS images only. However, DUFL aims to use generative networks to learn HR-HS images from a random noise and therefore does not take full use of the available information in the observations such as the HR-RGB image with high spatial resolution structures. Subsequently, Uezato et al. [25] used a deep decoder network to generate the latent HR-HS images from both noisy input data as well as the observed HR-RGB observations as guided information, called a guided deep decoder (GDD) network. In addition, Fu et al. [7] propose to conduct joint optimization of the optimal CSF and the potential HR-HS images from the observed observations only. Although these current unsupervised methods illustrate the potential for plausible HR-HS image generation, most of them do not fully exploit the high-resolution spatial structure of the observed HR-RGB images. Therefore, there is still room for improvement in terms of performance.

To handle the above mentioned issues, this study proposes a new deep RGB-driven generative network for unsupervised HSI SR that uses the observed HR-RGB images instead of a random noise as the network input. Specifically, by leveraging the observed HR-RGB images with high-resolution spatial structure as the input, we design an encoder-decoder based two-dimensional convolutional network to learn the latent HR-HS image, and then follow the specially designed convolutional layers to implement the spatial and spectral degradation procedure for obtaining the approximated LR-HS and HR-RGB images. Thus the loss functions for training the generative network can be formulated using the observed LR-HS and HR-RGB image only where no external samples are required in the end-to-end unsupervised learning model. Experimental results on two HS datasets show that our method outperforms the state-of-the-art methods. The main advantages of this study are summarized as follows.

I. We propose a RGB-driven generative network by making full use of the high-resolution spatial structure in the observed RGB image as the conditional input for robust HR-HS image estimation.

II. We learn the specific CNN model directly from the observations without the need for a labeled training set.

III. We construct a simple convolution-based degradation modules using the specifically designed depth-wise and point-wise convolution layers for imitating the observation procedure, which are easy to be optimized with generative networks.

2 Proposed Method

In this section, we first introduce the problem of formulating the HSI SR task and then describe our proposed deep RGB-driven generative network.

2.1 Problem formulation

The goal of the HSI SR task is to generate HR-HS images by combining the LR-HS and HR-RGB images: $\mathbf{I_y} \in \Re^{\mathbf{w}\times\mathbf{h}\times\mathbf{L}}$ and $\mathbf{I_x} \in \Re^{\mathbf{W}\times\mathbf{H}\times\mathbf{3}}$, where $W(w)$ and $H(h)$ denote the width and height of the HR-HS (LR-HS) image and $L(3)$ denotes the spectral number. In general, the degradation process of the observations: $\mathbf{I_x}$ and $\mathbf{I_y}$ from the HR-HS image $\mathbf{I_z}$ can be mathematically expressed as follows.

$$\mathbf{I_x} = \mathbf{k}^{(\mathbf{Spa})} \otimes \mathbf{I_z}^{(\mathbf{Spa})} \downarrow + \mathbf{n_x}, \mathbf{I_y} = \mathbf{I_z} * \mathbf{C}^{(\mathbf{Spec})} + \mathbf{n_y}, \tag{1}$$

where $\otimes$ denotes the convolution operator, $\mathbf{k}^{(\mathbf{Spa})}$ is the two-dimensional blur kernel in the spatial domain, and $(Spa)\downarrow$ denotes the down-sampling operator in the spatial domain. $\mathbf{C}^{(\mathbf{Spec})}$ is the spectral sensitivity function of the RGB camera (three filters in the one-dimensional spectral direction), which converts L spectral bands into RGB bands, and $\mathbf{n_x}$, $\mathbf{n_y}$ are additive white Gaussian noise (AWGN) with a noise level of σ. For simplicity, we rewrite the mathematical degradation model in Eq. 1 as the following matrix form.

$$\mathbf{I_x} = \mathbf{D}\mathbf{I_z} + \mathbf{n_x}, \mathbf{I_y} = \mathbf{I_z}\mathbf{C} + \mathbf{n_y}, \tag{2}$$

where $\mathbf{D}$ is the spatial degradation matrix containing the spatial blurring matrix and the down-sampling matrix, and $\mathbf{C}$ is the spectral transformation matrix representing the camera spectral sensitivity function (CSF). By assuming the known spatial and spectral degradations, the HSI SR problem can be solved intuitively through minimizing the following reconstruction errors.

$$\mathbf{I_z}^* = \arg\min_{\mathbf{I_z}} \|\mathbf{I_x} - \mathbf{D}\mathbf{I_z}\|_F^2 + \|\mathbf{I_y} - \mathbf{I_z}\mathbf{C}\|_F^2, \tag{3}$$

where $|\cdot|_F$ denotes the Frobenius norm. However, Eq. 3 may have several optimal solutions that yield minimal reconstruction errors, and thus direct optimization would lead to a very unstable solution. Most existing methods typically use different hand-crafted priors to model the potential HR-HS to impose the constraints on Eq. 3 for narrowing the solution space, and demonstrate great performance improvement with elaborated priors. The prior-regularized mathematical model approach is expressed as follows.

$$\mathbf{I_z}^* = \arg\min_{\mathbf{I_z}} \|\mathbf{I_x} - \mathbf{D}\mathbf{I_z}\|_F^2 + \|\mathbf{I_y} - \mathbf{I_z}\mathbf{C}\|_F^2 + \alpha\phi(\mathbf{I_z}), \tag{4}$$

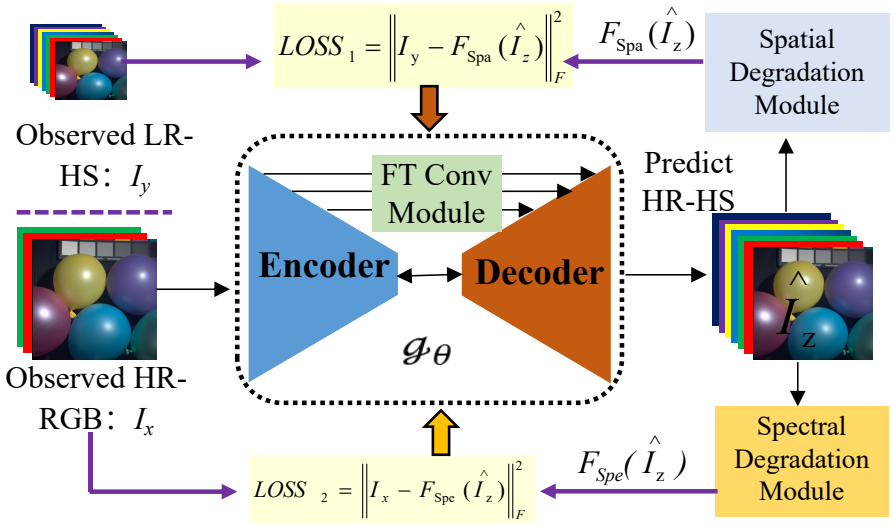

Fig. 1. Proposed framework of deep RGB-driven generative network. FT Conv Module denotes feature transfer convolution module. Spatial Degradation module is implemented by a specifically designed depth-wise convolution layer while Spectral Degradation module is realized by a point-wise convolution layer.

where $\phi(\mathbf{I_z})$ is used as the regularization term for modeling the prior in the latent HR-HS image, while α is the hyper-parameter for adjusting the contribution of the regularization term and the reconstruction error. However, the investigated priors in the existing methods are designed empirically and usually face difficulty to sufficiently modeling the complicated spatial and spectral structures.

2.2 Proposed deep RGB-driven generative network

As shown in many vision tasks, deep convolutional networks have powerful modeling capabilities to capture the inherent prior knowledge of different visual data (images), and in this study, deep learning networks are used to automatically learn the prior knowledge embedded in HR-HS images. Specifically, we use an encoder-decoder-based generative network to automatically reconstruct HR-HS images. In the absence of the ground-truth HR-HS images for training the generative network as in the conventional fully supervised learning networks, we employ the observed HR-RGB and LR-HS images to formulate the loss functions expressed in Eq. 3.

Specifically, given the predicted HR-HS image $\hat{\mathbf{I}}_\mathbf{z} = g_\theta(\cdot)$, where g_θ denotes the generative network and θ is its parameter, we specifically designed a depth-wise convolutional layer to implement a spatially degradation model as $F_{Spa}(\hat{\mathbf{I}}_\mathbf{z})$ and a point-wise convolutional layer to implement the spectral transform $F_{Spe}(\hat{\mathbf{I}}_\mathbf{z})$. By simply fixing the weights of the specially designed convolu-

tional layers in the spatial degradation matrix $\mathbf{D}$ and the CSF matrix $\mathbf{C}$, we can transform the output of the generative network into the approximated versions of the HR-RGB image : $\mathbf{I_x}$ and the LR-HS image: $\mathbf{I_y}$. According to Eq. 3, we minimize the reconstruction errors of the under-studying LR-HS and HR-RGB images to train the generative network, formulated as the follows:

$$\theta^* = \arg\min_{\theta} \|\mathbf{I}_x - F_{Spe}(g_\theta(\cdot))\|_F^2 + \|\mathbf{I}_y - F_{Spa}(g_\theta(\cdot))\|_F^2, \tag{5}$$

Since the generative network g_θ can potentially learn and model the inherent priors in the latent HR-HS image, it is not necessary to explicitly impose prior modeling constraints as the regularization termin the Eq. 5. The conceptual framework of the proposed deep RGB-driven generativenetwork is shown in Fig. 1. It can be trained using the observed LR-HS and HR-RGB image only without any external data. In the following subsections, we will present the generative network architecture and the network inputs.

The generative network architecture: For the generative network g_θ, any DCNN can be used to serve as the beseline narchitecture in our proposed framework. Since the latent HR-HS images often contain various structures, rich textures, and complex spectra, the employed generative network g_θ has to possess enough modeling ability to ensure reliable HR-HS image representation. Several generative architectures [2] have been investigated and significant progress has been made in generating high-quality natural images [14], for example in the context of the adversarial learning research. Since our unsupervised framework requires training a specific CNN model for each under-studying observation, shallower networks are preferred to reduce the training time. Moreover, it is known that a deeper network architecture, which can capture feature representation in a large receptive field can improve the super-resolution performance. Therefore, a shallow network with sufficient representation modeling capability in a larger receptive field would be suitable for our network structure.

It is well known that encoder-decoder networks have a shallow structure being possible to learn feature representation in large-scale spatial context due to down-sampling operations between adjacent scales, and thus we employ the encoder-decoder structure as our generative network. In detail, the generative network consists of an encoder subnet and a decoder subnet, and both encoder and decoder include multiple blocks with different scales that can capture feature at different receptive fields. The outputs of all blocks in the encoder subnet are transferred to the corresponding blocks in the decoder using a convolution-based feature transfer module (FT Conv module) to reuse the learned detailed features. Each block consists of three convolutional/RELU pairs, where a max pooling layer with 2×2 kernels are used to reduce the feature map size between the blocks of the encoder, and an up-sampling layer is used to double the feature map size between the blocks of the decoder. Finally, a vanilla convolution-based output layer is used to estimate the underlying HR-HS image.

The RGB-guided input: Most generative neural networks are trained to synthesize the target images with the specific defined concept from noisy vectors, which are randomly generated based on a distribution function (e.g., Gaussian

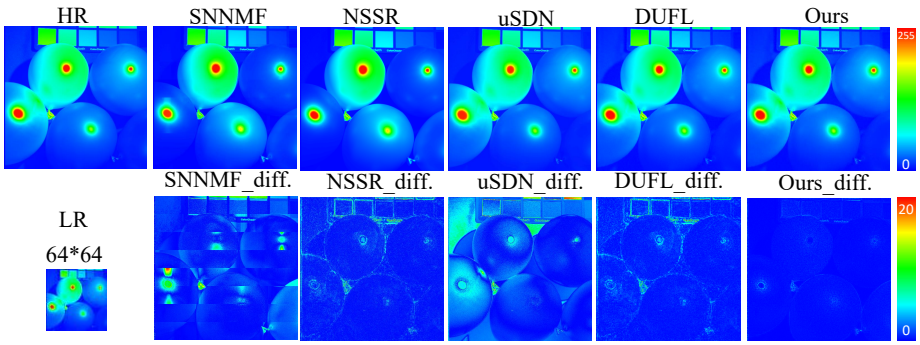

Fig. 2. Visual results of mathematical optimization-based methods: SNNMF [27], NSSR [6] and deep learning-based methods: uSDN [22], DUFL [18] on the CAVE dataset and our method with the up-scale factor 8.

or uniform distribution). As recent studies have confirmed, randomly generated noisy inputs usually produce sufficiently diverse and unique images. Our HSI SR task aims to use the observed LR-HS and HR-RGB images to learn the corresponding HR-HS images. Simply using noise as input does not take full advantage of the existed information in the observations. Therefore, we attempt to employ the available observation as a conditional guide for our generative network.

The observed HR-RGB images are known to have a high spatial resolution structure, and it is expected to assist the two-dimension convolution-based generative network learning more effective representation for reliable HR-HS image recovery. The observed LR-HS images can also be used as conditional inputs to the network. However, the low-resolution spatial structure may lead to local minimization of the network training process, which may have a negative impact on the prediction results. More importantly, the magnification factor, e.g., 10 for 31 spectral bands estimation from RGB in the spectral domain,, is usually much smaller than in the spatial domain (64 in total (8×8) with an up-sampling factor of 8), so we use the observed HR-RGB image as conditional network input, denoted as $\mathbf{Z}^* = g_\theta(\mathbf{I}_y)$.

3 Experiment Result

3.1 Experimental Setting

We evaluated our method on two commonly used datasets, Cave and Harvard datasets. The Cave dataset contains 32 HS images taken in real material and object space, all with the same spatial resolution, e.g. 512×512 with 31 adjacent spectral bands ranging from 400 nm to 700 nm. The Harvard dataset contains 50 HS images taken during daylight hours, both outdoors and indoors, all with the same spatial resolution of 1392×1040 and 31 spectral bands ranging from 420

Table 1. Compared results with the SoTA methods including mathematical optimization-based and deep learning-based methods on both CAVE and Harvard datasets with the up-scale factors 8 and 16.

		CAVE					Harvard				
		RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
		Up-scale factor =8									
Mathematical optimization	GOMP	5.69	33.64	-	11.86	2.99	3.79	38.89	-	4.00	1.65
	MF	2.34	41.83	-	3.88	1.26	1.83	43.74	-	2.66	0.87
	SNNMF	1.89	43.53	-	3.42	1.03	1.79	43.86	-	2.63	0.85
	CSU	2.56	40.74	0.985	5.44	1.45	1.40	46.86	0.993	1.77	0.77
	NSSR	1.45	45.72	0.992	2.98	0.80	1.56	45.03	0.993	2.48	0.84
Deep learning	SSFNet	1.89	44.41	0.991	3.31	0.89	2.18	41.93	0.991	4.38	0.98
	DHSIS	1.46	45.59	0.990	3.91	0.73	1.37	46.02	0.981	3.54	1.17
	ResNet	1.47	45.90	0.993	2.82	0.79	1.65	44.71	0.984	2.21	1.09
	uSDN	4.37	35.99	0.914	5.39	0.66	2.42	42.11	0.987	3.88	1.08
	DUFL	2.08	42.50	0.975	5.36	1.16	2.38	42.16	0.965	2.35	1.09
	Ours	1.35	46.20	0.992	3.05	0.77	1.07	49.17	0.994	1.59	0.72
	Up-scale factor = 16										
Mathematical optimization	GOMP	6.08	32.96	-	12.60	1.43	3.83	38.56	-	4.16	0.77
	MF	2.71	40.43	-	4.82	0.73	1.94	43.30	-	2.85	0.47
	SNNMF	2.45	42.21	-	4.61	0.66	1.93	43.31	-	2.85	0.45
	CSU	2.87	39.83	0.983	5.65	0.79	1.60	45.50	0.992	1.95	0.44
	NSSR	1.78	44.01	0.990	3.59	0.49	1.65	44.51	0.993	2.48	0.41
Deep learning	SSFNet	2.18	41.93	0.991	4.38	0.98	1.94	43.56	0.980	3.14	0.98
	DHSIS	2.36	41.63	0.987	4.30	0.49	1.87	43.49	0.983	2.88	0.54
	ResNet	1.93	43.57	0.991	3.58	0.51	1.83	44.05	0.984	2.37	0.59
	uSDN	3.60	37.08	0.969	6.19	0.41	9.31	39.39	0.931	4.65	1.72
	DUFL	2.61	40.71	0.967	6.62	0.70	2.81	40.77	0.953	3.01	0.75
	Ours	1.71	44.15	0.990	3.63	0.48	1.28	47.37	0.992	1.92	0.49

Table 2. Ablation studies of different numbers of employed blocks in the generative network and loss terms on CAVE dataset with the up-scale factor 8.

Number of employed blocks	Loss	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
2	Both	1.45	45.49	0.992	3.47	0.81
3		1.42	45.69	0.992	3.28	0.81
4		1.38	46.05	0.993	3.13	0.77
5	Loss1	26.27	19.85	0.601	43.53	16.19
	Loss2	3.30	38.57	0.972	3.68	1.88
	Both	1.35	46.20	0.992	3.05	0.77

Table 3. Ablation studies of different network inputs on CAVE dataset with the up-scale factor 8.

Block number:5 Loss: both					
Input	RMSE↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
Noise	2.10	42.53	0.978	5.30	1.12
Combined	1.46	45.47	0.992	3.27	0.81
Combined + noise	1.44	45.61	0.992	3.72	0.80
Ours (RGB)	1.35	46.20	0.992	3.05	0.77

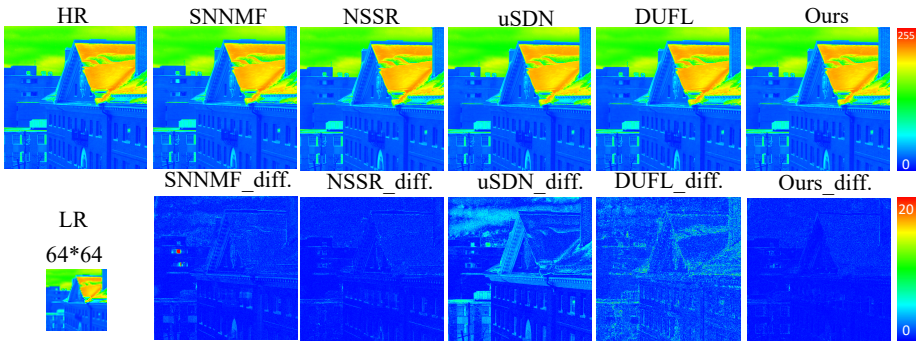

Fig. 3. Visual results of mathematical optimization-based methods: SNNMF [27], NSSR [6] and deep learning-based methods: uSDN [22], DUFL [18] on the Harvard dataset and our method with the up-scale factor 8.

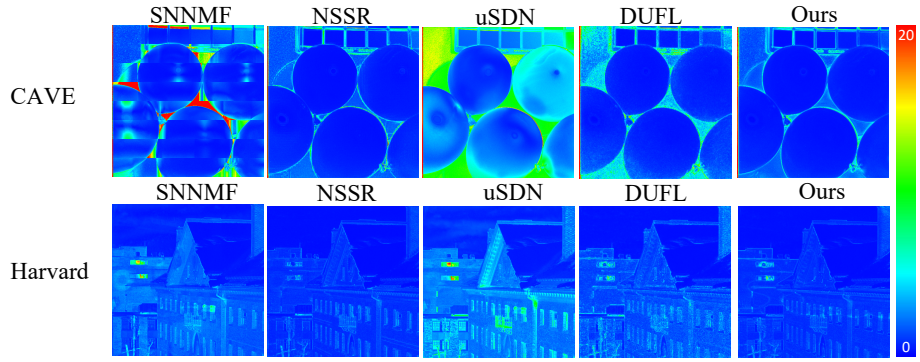

Fig. 4. SAM vuisual results of mathematical optimization-based methods: SNNMF [27], NSSR [6] and deep learning-based methods: uSDN [22], DUFL [18] on the CAVE and Harvard datasets and our method with the up-scale factor 8.

nm to 720 nm. For both datasets, we transformed the corresponding HS images using the spectral response function of the Nikon D700 camera to obtain HR-RGB HS images, while LR-HS images were obtained by bicubic downsampling of the HS images. To objectively evaluate the performance of different HSI SR methods, we adopted five widely used metrics, including root mean square error (RMSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM), spectral angle mapper (SAM), and relative dimensional global errors (ERGAS).

First, we conducted experiments using the generative network with five blocks, all two terms of loss in Eq. 5 and the RGB input for comparing with the state-of-the-art methods, and then we performed an ablation study by varying the number of blocks, loss terms, and inputs to the generative network.

3.2 Comparisons with the State-of-the-art Methods

We compared our approach with various state-of-the-art methods, including those based on mathematical optimization-based methods: GOMP [26], MF [15], SNNMF [27], CSU [31], NSSR [6], supervised deep learning-based methods: SSFNet [9], DHSIS [5], ResNet [11], and unsupervised deep learning-based methods: uSDN [22], DUFL [18]. Table 1 shows the comparative results for the spatial expanding factor 8. Table 1 demonstrates that our method is able to significantly improve the performance in term of all evaluation metrics. In addition, Fig. 2 and Fig. 3 show the visualization difference results of two representative images with different deep unsupervised learning methods. Fig. 4 illuminates the SAM visualization results on both CAVE and Harvard datsets. It also manifests that our proposed method provides small reconstruction errors in most spatial locations.

3.3 Ablation Study

We validate the performance effect by varying the block (scale) numbers of the generative network, the used reconstruction error term, and the network inputs. As mentioned above, we used an encoder-decoder structure where both encoder and decoder paths contain multiple blocks as our specific CNN model to extract multi-scale contexts in different receptive fields. To test the efficiency of the used multiple scales, we varied the block number from 2 to 5 and performed HR-HS image learning experiments. The comparative results are shown in the Table 2, where more blocks demonstrates the improvement in term of the performance, while the generative network achieves impressive result even with only two blocks. In addition, as described in Eq. 5, we use the reconstruction errors of both observed HR-RGB and LR-HS images (denoted as 'both' loss) as loss functions, and we further take one term only in Eq. 5 as the loss formulas used to train our generative the network, denoted as loss 1 and loss 2 for comparison. Table 2 illustrates the comparison results using different loss functions, which indicates that the proposed two loss terms perform much better.

Finally, we verified the effect of different inputs to the generative network. As mentioned above, it is popular in most generative networks to use randomly generated noisy inputs to synthesize different natural images. To make full use of the available data, we employed the observed HR-RGB image as the conditional input to guide the training of the proposed generative network. Without lack of generality, we also combined the HR-RGB image with the up-sampled LR-HS image together as the network input (marked as 'combined') and additionally disturb the combined input with a small level of noise in each training step to increase the robustness of model training. The comparison results of different network inputs are shown in Table 3, and the conditional input using the HR-RGB image manifests the best recovery performance.

4 Conclusion

In this study, we proposed a new deep RGB-driven generative network that learns the latent HR-HS image from its degraded observations without the need of any external data. To build an efficient and effective specific CNN model, we adopted an encoder-decoder-based generative network with a shallow structure but being able to perform multi-scale spatial context exploration in large receptive fields to learn high-representative feature of the latent HR-HS image with the conditioned HR-RGB image as a guide. Moreover, since the under-studying scene do not have the ground-truth HR-HS image, we specifically designed the convolution-based degradation modules to transform the predicted HR-HS image in the generative network, and then obtained the approximated observations to formulate the loss function for network training. Experimental results showed that our method significantly improves the performance over the SoTA methods.

Acknowledgements This research was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20K11867, and JSPS KAKENHI Grant Number JP12345678.

References

1. Akhtar, N., Shafait, F., Mian, A.: Sparse spatio-spectral representation for hyperspectral image super-resolution. In: European conference on computer vision. pp. 63–78. Springer (2014)
2. Bach, S.H., He, B., Ratner, A., Ré, C.: Learning the structure of generative models without labeled data. In: International Conference on Machine Learning. pp. 273–282. PMLR (2017)
3. Bioucas-Dias, J.M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N., Chanussot, J.: Hyperspectral remote sensing data analysis and future challenges. IEEE Geoscience and remote sensing magazine **1**(2), 6–36 (2013)
4. Dian, R., Fang, L., Li, S.: Hyperspectral image super-resolution via non-local sparse tensor factorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5344–5353 (2017)
5. Dian, R., Li, S., Guo, A., Fang, L.: Deep hyperspectral image sharpening. IEEE transactions on neural networks and learning systems (99), 1–11 (2018)
6. Dong, W., Fu, F., Shi, G., Cao, X., Wu, J., Li, G., Li, X.: Hyperspectral image super-resolution via non-negative structured sparse representation. IEEE Transactions on Image Processing **25**(5), 2337–2352 (2016)
7. Fu, Y., Zhang, T., Zheng, Y., Zhang, D., Huang, H.: Hyperspectral image super-resolution with optimized rgb guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11661–11670 (2019)
8. Han, X.H., Chen, Y.W.: Deep residual network of spectral and spatial fusion for hyperspectral image super-resolution. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). pp. 266–270. IEEE (2019)
9. Han, X.H., Shi, B., Zheng, Y.: Self-similarity constrained sparse representation for hyperspectral image super-resolution. IEEE Transactions on Image Processing **27**(11), 5625–5637 (2018)

10. Han, X.H., Shi, B., Zheng, Y.: Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 2506–2510. IEEE (2018)
11. Han, X.H., Sun, Y., Chen, Y.W.: Residual component estimating cnn for image super-resolution. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). pp. 443–447. IEEE (2019)
12. Han, X.H., Zheng, Y., Chen, Y.W.: Multi-level and multi-scale spatial and spectral fusion cnn for hyperspectral image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
13. He, W., Zhang, H., Zhang, L., Shen, H.: Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration. IEEE transactions on geoscience and remote sensing **54**(1), 178–188 (2015)
14. Huang, Q., Li, W., Hu, T., Tao, R.: Hyperspectral image super-resolution using generative adversarial network and residual learning. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3012–3016. IEEE (2019)
15. Kawakami, R., Matsushita, Y., Wright, J., Ben-Ezra, M., Tai, Y.W., Ikeuchi, K.: High-resolution hyperspectral imaging via matrix factorization. In: CVPR 2011. pp. 2329–2336. IEEE (2011)
16. Lanaras, C., Baltsavias, E., Schindler, K.: Hyperspectral super-resolution by coupled spectral unmixing. In: Proceedings of the IEEE international conference on computer vision. pp. 3586–3594 (2015)
17. Liu, Z., Zheng, Y., Han, X.H.: Unsupervised multispectral and hyperspectral image fusion with deep spatial and spectral priors. In: Proceedings of the Asian Conference on Computer Vision (2020)
18. Liu, Z., Zheng, Y., Han, X.H.: Deep unsupervised fusion learning for hyperspectral image super resolution. Sensors **21**(7), 2348 (2021)
19. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. Journal of biomedical optics **19**(1), 010901 (2014)
20. Mertens, S., Verbraeken, L., Sprenger, H., Demuynck, K., Maleux, K., Cannoot, B., De Block, J., Maere, S., Nelissen, H., Bonaventure, G., et al.: Proximal hyperspectral imaging detects diurnal and drought-induced changes in maize physiology. Frontiers in plant science **12**, 240 (2021)
21. Park, S.M., Kim, Y.L.: Spectral super-resolution spectroscopy for biomedical applications. In: Advanced Chemical Microscopy for Life Science and Translational Medicine 2021. vol. 11656, p. 116560N. International Society for Optics and Photonics (2021)
22. Qu, Y., Qi, H., Kwan, C.: Unsupervised sparse dirichlet-net for hyperspectral image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2511–2520 (2018)
23. Sakthivel, S.P., Sivalingam, J.V., Shanmugam, S., et al.: Super-resolution mapping of hyperspectral images for estimating the water-spread area of peechi reservoir, southern india. Journal of Applied Remote Sensing **8**(1), 083510 (2014)
24. Saralıoğlu, E., Görmüş, E.T., Güngör, O.: Mineral exploration with hyperspectral image fusion. In: 2016 24th Signal Processing and Communication Application Conference (SIU). pp. 1281–1284. IEEE (2016)
25. Uezato, T., Hong, D., Yokoya, N., He, W.: Guided deep decoder: Unsupervised image pair fusion. In: European Conference on Computer Vision. pp. 87–102. Springer (2020)
26. Wang, J., Kwon, S., Shim, B.: Generalized orthogonal matching pursuit. IEEE Transactions on signal processing **60**(12), 6202–6216 (2012)

27. Wycoff, E., Chan, T.H., Jia, K., Ma, W.K., Ma, Y.: A non-negative sparse promoting algorithm for high resolution hyperspectral imaging. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1409–1413. IEEE (2013)
28. Xie, Q., Zhou, M., Zhao, Q., Meng, D., Zuo, W., Xu, Z.: Multispectral and hyperspectral image fusion by ms/hs fusion net. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1585–1594 (2019)
29. Xu, J.L., Riccioli, C., Sun, D.W.: Comparison of hyperspectral imaging and computer vision for automatic differentiation of organically and conventionally farmed salmon. Journal of Food Engineering **196**, 170–182 (2017)
30. Yokoya, N., Chan, J.C.W., Segl, K.: Potential of resolution-enhanced hyperspectral data for mineral mapping using simulated enmap and sentinel-2 images. Remote Sensing **8**(3), 172 (2016)
31. Yokoya, N., Zhu, X.X., Plaza, A.: Multisensor coupled spectral unmixing for time-series analysis. IEEE Transactions on Geoscience and Remote Sensing **55**(5), 2842–2857 (2017)
32. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L.: Image super-resolution: The techniques, applications, and future. Signal Processing **128**, 389–408 (2016)
33. Zhang, S., Liang, G., Pan, S., Zheng, L.: A fast medical image super resolution method based on deep learning network. IEEE Access **7**, 12319–12327 (2018)
34. Zhao, Y., Yang, J., Zhang, Q., Song, L., Cheng, Y., Pan, Q.: Hyperspectral imagery super-resolution by sparse representation and spectral regularization. EURASIP Journal on Advances in Signal Processing **2011**(1), 1–10 (2011)
35. Zhu, Z., Hou, J., Chen, J., Zeng, H., Zhou, J.: Residual component estimating cnn for image super-resolution. vol. 30, pp. 1423–1428 (2020)