Symmetrical Feature Propagation Network for Hyperspectral Image Super-Resolution

Qiang Li, Maoguo Gong, *Senior Member, IEEE,* Yuan Yuan, *Senior Member, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Single hyperspectral image (HSI) super-resolution (SR) methods using a auxiliary high resolution RGB image have achieved great progress recently. However, most existing methods aggregate the information of RGB image and HSI early during input or shallow feature extraction, whose the difference between two images has not been treated and discussed. Although a few methods combine both image features in the middle layer of the network, they fail to make full use of the two inherent properties, i.e., rich spectra of HSI and HR content of RGB image, to guide model representation learning. To address these issues, in this paper, we propose a dual-stage learning approach for HSI SR to learn a general spatial-spectral prior and image-specific details, respectively. In coarse stage, we fully take advantage of two adjacent bands and RGB image to build model. During coarse SR, a symmetrical feature propagation approach is developed to learn the inherent content of each image over a relatively long range. The symmetrical structure encourages the two streams to better retain their particularity. Meanwhile, it can realize the information interaction by adaptive local block aggregation module. To learn the image-specific details, a back-projection refinement network is embedded in the structure, which further improves the performance in fine stage. The experiments on four benchmark datasets demonstrate that the proposed approach presents excellent performance over the existing methods. Our code is publicly available at https://github.com/qianngli/SFPN.**

***Index Terms*—Hyperspectral image, super-resolution, symmetrical structure, separation-and-aggregation, block aggregation.**

I. Introduction

HYPERSPECTAL image (HSI) contains dozens or even hundreds of continuous bands collected by spectral imaging system within a certain range of the spectrum, which can accurately depict more faithful knowledge in real scenes [1]. Since HSI provides rich spectra compared to natural image or multispectral image, it is more helpful for representation learning [2]. Benefiting from this characterization, various natural image tasks greatly improve performance by combining HSI, such as object recognition [3], [4], object tracking [5], etc.

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, U1864204, and 61825603.

Qiang Li is with the School of Electronic Engineering, Xidian University, Xi'an 710071, P.R. China (e-mail: liqmges@gmail.com).

Maoguo Gong is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an 710071, P.R. China (e-mail: gong@ieee.org).

Yuan Yuan and Qi Wang are with the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: y.yuan1.ieee@gmail.com, crabwq@gmail.com) (*Corresponding author: Qi Wang.*)

(a) RGB image (b) band on 530 nm (c) band on 610 nm

Fig. 1. Illustration of texture difference both in bands and RGB image.

In real cases, optical imaging often faces a tradeoff between spatial and spectral resolution. For instance, a natural image exhibits high spatial resolution, yet its spectral resolution is low. On the contrary, HSI yields better spectral resolution, whereas it has lower spatial resolution [6], [7]. In some scenes, images with both high spectral and spatial resolution are desirable, so that better feature representation can be obtained. A natural way is to fuse HSI with low spatial resolution and RGB image with high spatial resolution to generate high resolution (HR) HSI in spatial and spectral domain simultaneously, which is known as HSI super-resolution (SR).

Recent studies have shown various works for SR by merging low resolution (LR) HSI and HR RGB image, including traditional methods [8]–[11] and deep learning-based methods [12]–[15]. Here, resolution refers to the spatial resolution. Existing traditional methods have handcrafted diverse shallow prior knowledge, e.g., sparsity [16] and low-rank [17]. These techniques mainly exploit such as Bayesian [18], matrix factorization [19] to encode this knowledge into the optimization model. After the constraint conditions are set, the super-resolved HSI is obtained via multiple iterations and optimizations. In challenging situations, as the feature representation of traditional approaches is limited, it can not be popularized well.

Deep learning-based approaches have exhibited superior performance in natural image SR task. Recent works address HSI SR by referring to SR algorithms for natural image [20]–[22], including unsupervised and supervised manner. Compared with traditional methods, these studies require less prior knowledge or none. Benefiting from the strong representation ability of convolutional neural network, unsupervised fusion approaches [23]–[25] have made remarkable progress. However, these methods exists in two issues. The one is that some methods require multiple iterations when test, such as [23], which increase the execution time. The other is that this type of methods obtain poor performance for real LR HSI, compared with supervised deep learning-based methods. Therefore, we focus on constructing model using supervised manner in our

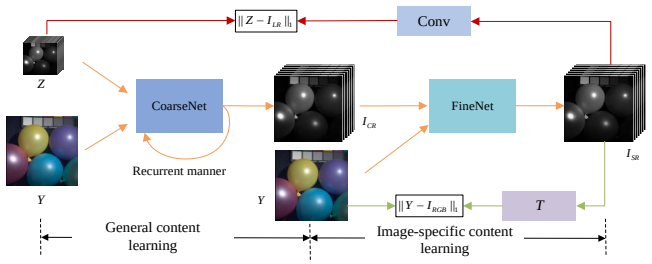

Fig. 2. The proposed coarse-to-fine learning structure for HSI SR.

paper.

Although public HSI dataset contains a small number of training samples, the researchers augment these data to train the model in a supervised manner [26], [27]. Typical models contain MHF-net [28], PZRes-Net [27], etc. In contrast, supervised fusion methods obtain better performance. Nevertheless, there are still some obvious shortcomings. For instance, most existing methods simply input HSI and HR RGB image into the model through the stack or in parallel, and integrate them at the beginning or initial feature extraction. Actually, edge texture exists obvious distinction between RGB image and HSI, which is shown in Fig. 1. This unites the two images prematurely, so that the difference between them has not been treated and discussed. While a few methods [15], [29] combine both features in the middle layer of the network, they do not make effective use of significant properties, i.e., rich spectral information of HSI and HR content of RGB image. Concretely, HSI exhibits a remarkable characteristic that the adjacent bands have similarity. These adjacent bands help to complement each other. Besides, as seen in Fig 1, RGB image has sharper detail content than some bands. Ideally, the combination of these two properties is more favorable for the exploration of HSI SR, which makes the super-resolved image get clear details. Therefore, how to distinguish the differences between them and effectively make use of their significant properties is still an urgent problem to be solved.

Motivated by the above observations, in this paper, we propose a coarse-to-fine learning approach for HSI SR, which is shown in Fig. 2. The framework mainly consists of CoarseNet and FineNet. The two networks here learn a general spatial-spectral prior and image-specific details, respectively. Specifically, a symmetrical feature propagation model is designed to general spatial-spectral prior in coarse stage, i.e., CoarseNet. Encouraged by symmetrical multi-step propagation mechanism, the initial results are produced in a recurrent way through the effective combination of multiple bands and HR RGB image. On this basis, a back-projection refinement network, i.e., FineNet, is added to the end of CoarseNet to learn the image-specific details by cycle consistency loss, which further improves the performance of model. Experiments demonstrate that the proposed method exhibits state-of-the-art results. In summary, the contributions of this paper are three-fold:

- A symmetrical multi-step strategy is developed to explore the information of two modalities, forming a symmetrical structure. Different from existing methods, it can retain the inherent knowledge of the two modalities for a large range. Meanwhile, it can realize the information interaction in a local range.
- An adaptive local block aggregation module is proposed to achieve feature enhancement. The relationship between a single block and local blocks in its neighborhood is dynamically constructed by using unprocessed HR RGB image. According to the constructed relationship, the feature blocks with high similarity are aggregated together to realize the mining of local context information.
- The proposed network is evaluated and compared with mainstream methods. Extensive experiments demonstrate the effectiveness of adaptive local block aggregation module. Moreover, the performance of our work outperforms the competitors in real and synthetic datasets, resulting in high spectral fidelity and clear texture.

The rest of this paper is organized as follows: Section II reviews the related works based on fusion for HSI SR. Section III describes the proposed method in detail, including symmetrical feature propagation network and back-projection refinement network. The experiments are analyzed and discussed in Section IV. Finally, this work is concluded in Section V.

II. Realted Work

Currently, fusion-based SR approaches are mainly divided into two categories, including traditional methods and deep learning-based methods. While traditional algorithms have an extensive history, in this section, we only review deep learning-based methods, including unsupervised and supervised manner.

A. Unsupervised Methods

Although deep learning has achieved great success in various applications over the past few years, only a few works have been devoted to unsupervised fusion. For instance, Qu *et al.* [23] attempt first time to solve the HSI SR problem by unsupervised learning. Wang *et al.* [24] develop a variational probabilistic model to jointly optimize three subnetworks. Inspired by the deep network itself carrying a large amount of low-level prior knowledge, Liu *et al.* [30] construct a simple network to learn spatial and spectral prior. Uezato *et al.* [25] state that previous studies apply specific prior content and not design general model to address existing problems. To address this issue, a guided deep encoder network is proposed as a general prior knowledge to deal with image fusion in an unsupervised way. Although the above methods obtain better performance, some methods require multiple iterations, such as [23], which increase training time. Besides, the above methods obtain poor performance for real LR HSI, compared with supervised deep learning-based methods. Considering these problems, we turn our attention to constructing model using supervised manner in our paper.

B. Supervised Methods

According to input mode, existing supervised methods can be roughly divided into two categories, i.e., stack and parallel input mode.

1) Stack Input Mode: For stack input mode, this type of algorithm first upsamples LR HSI with the same spatial resolution as the HR RGB image, and inputs both them into the model by stacking. For example, Han *et al.* [31] propose a spatial and spectral fusion framework. It concatenates HR RGB image and upsampled LR HSI into a cube as input to jointly explore spectral attributes and rich spatial content. Similarly, Zhang *et al.* [32] adopt the same way to build input data. Unlike above works, Hu *et al.* [33] first downsample HR RGB image to the same size as LR HSI in spatial resolution, then combine them as a whole and upsample using PixelShuffle operation [34]. Since edge texture exists obvious distinction between RGB image and HSI, which is shown in Fig. 1. This unites the two images prematurely, so that the difference between them has not been analyzed and discussed. At present, it is almost abandoned to build models by stack mode.

2) Parallel Input Mode: Currently, most existing HSI SR methods adopt parallel input mode to construct the model, i.e., LR HSI and HR RGB image are fed into the model in parallel [23], [27], [28]. For instance, Zhu *et al.* [27] present a progressive zero-centric residual network (PZRes-Net) by attaching LR HSI of different dimensions along spectral direction. Its aim is to learn a zero-centric residual content from both inputs. Wang *et al.* [35] develop a dense fusion framework for HSI SR, including measurement module and fusion module. The modules learn observation model and extract spatial-spectral information by iterative manner, respectively. MoG-DCN [36] takes a similar approach. The significant difference is that the HR RGB image without feature extraction is fused with the image generated by the observation model. Meanwhile, the obtained observation image is also merged with LR HSI after a series of operations. These above approaches effectively explore the difference of individual image before fusion. However, these algorithms cannot learn the inherent properties of each image over a relatively long range.

To study their specificity over a long range, Han *et al.* [29] propose a multi-scale spatial-spectral fusion architecture by two pathways with diverse directions. This manner can keep spatial and spectral attributes for two images. However, it can easily lead to fusion ambiguity, since the depth information is different in the fusion stage. Later, Zhang *et al.* [15] design a supervised image fusion network and an unsupervised adaptation learning network. The networks yield a general image prior and a specific super-resolved HSI, respectively. In this method, it does not identify the differential information between the two very well. Importantly, this method does not make full use of the two inherent properties (i.e., rich spectra of HSI and HR content of RGB image) to construct the model. Inspired by this approach with two stages, in our paper, we also adopt this manner to study SR task. In contrast, our work in coarse stage skillfully exploits the adjacent bands and RGB image to restore the current band without inputting all the bands. During SR, it can not only realize the exchange of information between data, but also guarantee its inherent properties over a long range.

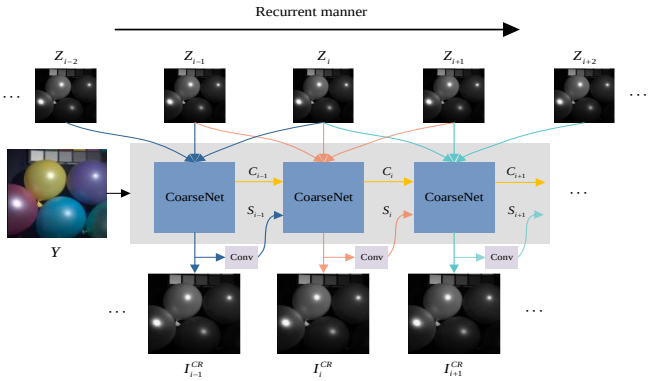

Fig. 3. Overview of flowchart in the coarse stage. Only three bands (current band Z_i and its two adjacent bands Z_{i-1}, Z_{i+1}) and RGB image are fed into CoarseNet to produce the initial super-resolved band I_i^{CR}. In this process, feature content C_{i-1} and spectral content S_{i-1} is injected into CoarseNet, which encourages the model to fuse complementary information, so as to improve the ability of feature representation.

III. Methodology

A. Problem Formulation

Given the two images, LR HSI $Z \in \mathbb{R}^{w\times h\times L}$ and HR RGB image $Y \in \mathbb{R}^{W\times H\times 3}$, the aim of SR task is to estimate HR HSI $X \in \mathbb{R}^{W\times H\times L}$ with both high spatial and spectral resolution using them. Here, w and h are the width and height of each band in LR HSI, respectively, and L denotes the total number of bands. W and H represent the width and height of each channel in the RGB image. In general, $w \ll W$ and $h \ll H$. The relation among three images can be formulated as

$$Y = XT, \tag{1}$$

$$Z = RX, \tag{2}$$

where $T \in \mathbb{R}^{L\times 3}$ denotes the spectral response function to transfer the dozens spectra into three channels, and $R \in \mathbb{R}^{wh\times WH}$ is a degradation function from X to Z.

B. Motivation and Overview

LR HSI exhibits low spatial resolution, and its spectral resolution is actually still high. The main purpose of HSI SR is to improve spatial resolution, leading to both high spatial and spectral resolution. Therefore, we should focus more on spatial study when establishing models. Besides, different from the RGB image, HSI has dozens or even hundreds of bands. When the spatial resolution and batch size of both images are the same, the HSI requires more memory footprint after the input model than the RGB image. Under limited hardware resources, this puts forward higher requirements for model construction with more layers. One natural way is to reduce batch size, but the training time will be greatly increased, making the parameter correction more slowly. When the batch size is set too small, the gradient descent direction of the model is not accurate, and the training results are easy to produce large shocks. This makes it difficult for the model to converge. Hence, the crucial issue is how to strengthen the exploration of spatial information while reducing memory footprint.

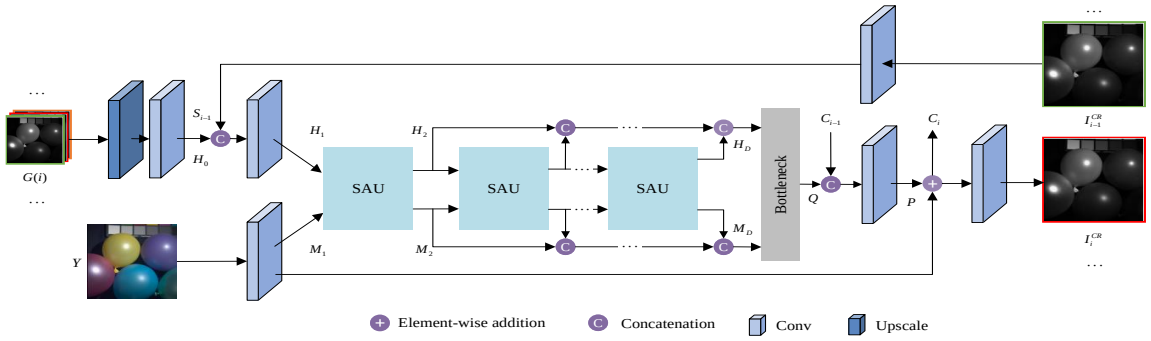

Fig. 4. The structure of proposed CoarseNet.

Considering this condition, we propose a symmetrical feature propagation network in coarse stage, which is shown in Fig. 3. Specifically, inspired by the high correlation among neighboring bands [37], we employ several bands and HR RGB image instead of inputting the whole HSI to achieve SR in coarse stage. By doing so, it promotes the analysis of the spatial resolution of the current band. Besides, this approach alleviates memory footprint only using several bands. To better apply adjacent spectral bands with high similarity, in our work, the current band and its two adjacent bands in the LR HSI are taken out separately and input into the network together with RGB image. It fully takes advantage of two inherent properties, namely rich spectra of HSI and HR content of RGB image. Through the fusion in space between them, we obtain current super-resolved band. All the bands are produced in the above way in a recurrent manner. During SR, the symmetrical multi-step propagation mechanism is designed to make the model get better feature representation. It not only retains their particularity, but also achieves the information interaction by means of separation-and-aggregation unit.

In the coarse stage, we only turn to the neighboring bands and ignore the relatively distant bands. Actually, these bands display distinct textures at diverse wavelengths (see Fig. 1). If they are handled effectively, it enables the model to yield clearer edges. Additionally, CoarseNet only generates the general structure for LR HSI, however fails to describe image-specific details, such as LR HSI with unknown degeneration. Considering that there is a fixed transformation T between initial result and RGB image, a back-projection refinement network is injected at the end of the structure to learn the image-specific details by cycle consistency loss, which can further refine the result. Note that all initial bands and RGB image are simultaneously fed into the network in fine stage.

C. Symmetrical Feature Propagation Network

To enhance the exploration of spatial information while reducing memory footprint, in our work, we develop a symmetrical feature propagation network for HSI SR. The overview of the flowchart is shown in Fig. 4. In this model, it mainly contains three parts, i.e., feature extraction, separation-and-aggregation unit (SAU), and feature fusion reconstruction.

1) Feature Extraction: Given the current band Z_i that needs to be restored, we exploit three bands $G(i)$ from HSI Z and corresponding image Y to conduct SR task. Here, three bands $G(i)$ are represented as

$$G(i) = \begin{cases} [Z_1, Z_2, Z_3], & i = 1 \\ [Z_{i-1}, Z_i, Z_{i+1}], & 2 \le i \le L-1. \\ [Z_{L-2}, Z_{L-1}, Z_L], & i = L \end{cases} \tag{3}$$

To diminish the increased parameters caused by multiple upsampling and downsampling, these bands $G(i)$ are upsampled to the same size as image Y using Nearest interpolation. Then, two branches are designed to study the respective features of $G(i)$ and Y using several convolution layers. Since the content between adjacent bands is highly similar, the information in the image is reinforced by combining them, which promotes further analysis of the image. Thus, the previous super-resolved band Z_{i-1} is introduced to help the current band Z_i to be reconstructed. Note that when restoring the first band Z_1, feature context does not participate in the fusion with shallow features, while it is all involved in the fusion for other bands $Z_i (i = 2, 3, ..., L)$, i.e.,

$$H_1(i) = \begin{cases} H_0, & i = 1 \\ f_{3\times3}[I_{i-1}^{CR}, H_0], & 2 \le i \le L \end{cases}, \tag{4}$$

where $f_{3\times3}(\cdot)$ represents convolution operation with kernel 3×3.

2) Separation-and-Aggregation Unit: Most previous works early integrate RGB image and LR HSI at the beginning or initial feature extraction. This mechanism cannot guarantee these two inherent forms in the SR process. Considering this point, a separation-and-aggregation unit (SAU) is developed to promote the feature fusion between the two forms of data, which is shown in Fig. 5. Meanwhile, the unit can preserve their particularity in the process of propagation through separation.

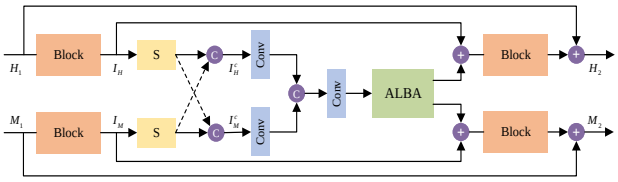

Fig. 5. Separation-and-aggregation unit (SAU).

Separation Unit: Let shallow features be H_1 and M_1 for two branches. They are input the first SAU simultaneously. To simply describe, we take one of the branches as an example.

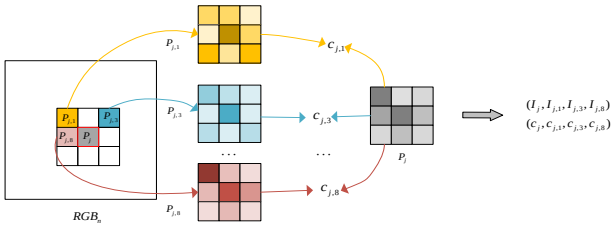

Fig. 6. The process of generating the position I and similarity c for block P_j.

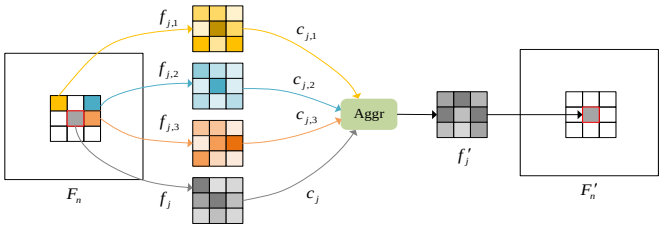

Fig. 7. The process of aggregating features among current block and three high similar blocks.

Concretely, the obtained features H_1 are explored in the block by

$$I_H = f_{3\times3}(ReLU(f_{3\times3}(H_1))), \tag{5}$$

where $ReLU(\cdot)$ is rectified linear unit activation function. Since the materials between the two branches are complementary to each other, to fully exploit this attribute, a natural approach is to combine complementary features in a certain position according to their representation ability. To achieve this end, we combine half of the feature maps of each branch $I_H, I_M \in \mathbb{R}^{C\times W\times H}$ to form new feature maps, i.e.,

$$I_H^c = [I_H^1, I_M^1], \tag{6}$$

$$I_M^c = [I_H^2, I_M^2], \tag{7}$$

where $I_H^1, I_H^2 \in \mathbb{R}^{(C/2)\times W\times H}$ and $I_M^1, I_M^2 \in \mathbb{R}^{(C/2)\times W\times H}$. C is the number of channels. This realizes the information exchange between the data and promotes the knowledge complementarity at the feature level. Subsequently, each branch learns the fused content by means of convolution layers with kernel 3×3, and the information of the two branches is stacked along the channel dimension.

Aggregation Unit: In the whole image, the object sometimes occupies a large area. If a small convolution kernel is utilized, it will ignore the exploration of adjacent content, so that the model cannot fully mine similar information. Although dilated convolution can increase the receptive field and improve the feature learning of the model in the local context, the corresponding relation between similar contents has not been established. To address this issue, in our paper, an adaptive local block aggregation (ALBA) module is designed. The main idea is to establish the relation between a single block and the local blocks in its neighborhood. Then, the feature blocks with high similarity are aggregated together to realize the mining of context information.

Since RGB image without feature extraction has HR content, it can accurately construct relation between blocks in local, including position index I and similarity c between the current block and its adjacent blocks. Specifically, given input HR RGB image Y, it is divided into small blocks with the same size $p \times p$. Without considering the edge blocks, cosine similarity is introduced to measure similarity between current block P_j and adjacent blocks $P_{j,k}$, i.e.,

$$c_{j,k} = \frac{< P_j, P_{j,k} >}{\|P_j\|_2 \times \|P_{j,k}\|_2}, k = 1, 2, \ldots, 8, \tag{8}$$

where $< \cdot, \cdot >$ denotes dot product operation, and $||\cdot||_2$ is L2 norm. Since the current block only has a certain similarity with a small number of blocks in the neighborhood, we sort the similarity values, and select first m similar blocks to generate the corresponding position index I. Fig. 6 shows an example of selecting three similar blocks. In particular, as for the blocks P_j at the four corners of the image Y, its three adjacent blocks are all involved in the calculation. Among the remaining blocks located at the edge in the image Y, only five adjacent blocks participated in the analysis.

After obtaining position index I and similarity c, we need to fuse depth features by block aggregation module. Likewise, the intermediate feature maps $F_n \in \mathbb{R}^{W\times H}$ are partitioned into small blocks $p \times p$ via the same strategy. Here, $n \in [1, ..., N]$ and N denotes batch size. The highly similar blocks $f_{j,1}, ..., f_{j,m}$ and the current block f_j are concatenated along the channel dimension by means of constructed position index I and similarity c. As a result, a new block $f_j^{'}$ is formed through a convolution layer with kernel 3×3, i.e.,

$$f_j^{'} = f_{3\times3}([f_j, f_{j,1} * c_{j,1}, ..., f_{j,m} * c_{j,m}]). \tag{9}$$

Fig. 7 displays an example of aggregating features. The proposed ALBA module effectively utilizes the local context information, thus enhancing the learning of similar content.

Since RGB image has sharper detail content than some bands, to distinguish the differences between them and effectively make use of their significant properties during SR, the features of the two modalities are separated after information fusion, and the corresponding branches are learned separately by block. It forms a symmetrical structure, which explicitly recalibrates the respective properties. The results after aggregation and separation are propagated to the next unit in the encoder, which forms multi-step propagation pattern. The mechanism encourages the two streams to better retain their particularity over a relatively long range, which is conducive to more accurate and effective coding of the two modalities.

3) Feature Fusion Reconstruction: After obtaining the deep features of multiple levels, the two modal features are gathered at the bottleneck, i.e.,

$$U = f_{1\times1}([H_2, ..., H_D]), \tag{10}$$

$$V = f_{1\times1}([M_2, ..., M_D]), \tag{11}$$

$$Q = f_{3\times3}([U, V]), \tag{12}$$

where $f_{1\times1}(\cdot)$ denotes convolution operation with kernel 1×1. Inspired by [38], we also introduce the feature context, i.e., the features C_{i-1} produced from the previous band SR are transferred back to the network for current band Z_i, which is denoted as

$$P = f_{1\times1}([Q, C_{i-1}]). \tag{13}$$

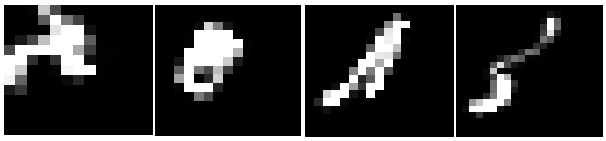

Fig. 8. Visual results of different blur kernels.

Finally, the model achieves estimated single band I_i^{CR} after convolution layer with kernel 3×3 and residual connection.

Coarse Network Training: To establish a more real training pair, five kernels with multiple sizes are randomly generated by using anisotropic and isotropic Gaussian kernels, including 7×7, 9×9, 13×13, 15×15 and 21×21. For anisotropic Gaussian kernel, the range of rotation angle is set to [0, π], and the range of kernel width is set to [0.2, s]. With respect to isotropic Gaussian kernel, the range of kernel width is set to [0.2, s]. According to the obtained blur kernel, two downsampling approaches are adopted to produce downsampled HSI. One is to directly convolute image with blur kernel. The other is to first blur image using blur kernel, and then downsampling by Bicubic or Bilinear interpolation. Finally, Gaussian noise is randomly added to degrade the LR HSI to reasonably represent the noise in the real scene. Through procedures mentioned above, we first synthesize large paired data $\{X, Z\}$. Furthermore, we train the coarse network by minimizing the loss $\mathcal{L}_{\mathcal{C}}$, i.e.,

$$\mathcal{L}_{\mathcal{C}}(\theta, i) = \min ||I_i^{CR} - X_i||_1, \tag{14}$$

where

$$I_i^{CR} = \mathcal{C}(s, Y, G(i); \theta), i = 1, 2, ..., L. \tag{15}$$

Here, $I_i^{CR} \in \mathbb{R}^{W \times H}$ and $|| \cdot ||_1$ is L1 nrom. The detailed process of training for coarse stage is described in **Algorithm 1**.

Algorithm 1: Training steps for coarse stage

Input: HR dataset $\mathcal{D}$ and scale factor s
Output: Coarse model parameter $\theta_{\mathcal{C}}$

```
Randomly initialize coarse model parameter θ;
while not done do
    Sample a btach of images {X} from dataset D;
    Generate {X, Z} by degradation, and obtain
      corrsponding HR image Y using Eq. 1;
    Set i = 1;
    for i≤L do
        Evaluate L_C by Eq. 14;
        Update θ according to L_C;
        i ← i + 1;
    end
end
```

D. Back-projection Refinement Network

In coarse stage, three bands are employed to perform SR task. The model only focuses on the adjacent bands and ignores the relatively distant bands. Additionally, CoarseNet only generates the general structure for LR HSI, however cannot effectively describe image-specific details, such as LR HSI with unknown degeneration. To capture more spectral knowledge and learn image-specific details, these coarse super-resolved bands $I_i^{CR} \in \mathbb{R}^{W \times H}, i = 1, 2, ..., L$ are combined as $I^{CR} \in \mathbb{R}^{W \times H \times L}$. Subsequently, we build the network with auxiliary RGB image Y and initial results I^{CR}, which is different from I^{CR} as the input alone in [15]. Concretely, a back-projection refinement network is embedded at the end of the coarse stage, whose structure is exhibited in Table I. To reduce the possible mapping functions, we believe that the content between LR HSI I and degraded super-resolved I^{SR} should be as consistent as possible. Motivated by [39], a cycle consistency loss is introduced to encourage this behavior, i.e.,

$$\mathcal{L}_{\mathcal{F}}(\sigma) = \min ||Y - I^{RGB}||_1 + \lambda ||Z - I^{LR}||_1, \tag{16}$$

$$I^{RGB} = \mathcal{F}(s, Y, I^{CR}; \sigma)T, \tag{17}$$

$$I^{LR} = \mathcal{B}(s, I^{SR}; \sigma), \tag{18}$$

where λ is a balanced factor. In our paper, λ is fixed as 0.1. With respect to the test process of two stages, it is elaborated in **Algorithm 2**.

Algorithm 2: Dual-stage test algorithm for HSI SR

Input: Given image I, trained coarse model parameter $\theta_{\mathcal{C}}$, scale factor s, and number of iterations Q
Output: Super-resolved image I^{SR}

```
Obtain corresponding RGB image Y by Eq. 1;
Generate LR image Z by downsampling I using
  different blur kernels (see Fig. 8), and add Gaussian
  noise;
Load coarse model parameter θ with θ_C;
I^CR = C(s, Y, G(i); θ_c);
Randomly initialize fine model parameter σ, and set
  q = 1;
for q ≤ Q do
    Evaluate L_F by Eq. 16;
    Update σ according to L_F;
    q ← q + 1;
end
return I^SR=F(s, Y, I^CR; σ_F)
```

TABLE I
THE STRUCTURE OF BACK-PROJECTION REFINEMENT NETWORK.

Type	Kernel size	Stride	Channel
Concat	–	–	34
Conv1	3×3	1	128
Conv2	3×3	1	128
Conv3	3×3	1	128
Conv4	3×3	1	128
Conv5	32×32	s	31

IV. Experiments

A. Datasets

1) CAVE: The dataset[1] was collected by cooled CCD camera, including wide variety of real world materials and objects [40]. It consists of 32 images, and each image has 31 bands with a spectral range of 400 nm to 700 nm at 10 nm step, where the resolution of band is 512×512 pixels. In our paper, 80% images are randomly selected as training set and the rest as test set.

2) Harvard: The dataset[2] was captured by Nuance FX, CRI Inc. under daylight illumination [41]. It contains 50 images from outdoor and indoors in the spectral range is from 420 nm to 720 nm at 10 nm step. The resolution of each image is 1392×1040 pixels. Similarly, we divide this dataset according to the above way for training and testing.

3) Chikusei: The dataset[3] was taken by Headwall Hyperspec-VNIR-C imaging sensor in Chikusei, Ibaraki, Japan. The HSI has 128 bands in the spectral range from 363 nm to 1018 nm, where each band contains 2517×2335 pixels. Unlike above datasets, it only has an image. We crop the top left of the HSI ($2000 \times 2335 \times 128$) as the training set, and other content of the HSI as the test set.

4) Sample of Roman Colosseum: The dataset[4] is a real remote sensing image obtained by WorldView-2. The dataset consists of an HR RGB image with size $1676 \times 2632 \times 3$ and an LR HSI with size $419 \times 658 \times 8$. Similarly, we select the top left of the LR HSI ($209 \times 658 \times 8$) and the corrsponding part of HR RGB image ($836 \times 2632 \times 3$) to train, and the remaining part of the dataset is exploited to test.

B. Comparison Methods and Evaluation Metrics

To demonstrate the superiority of proposed method, five approaches for HSI SR are employed, including LTTR [10], CMS [8], PZRes-Net [27], MHF-net [28], and UAL [15]. According to training manners, these methods can be divided into two parts, i.e., supervised and unsupervised, in which the supervised methods are LTTR and CMS, and the rest are unsupervised. Note that LTTR and CMS need to utilize the spectral response function many times in the process of SR.

To evaluate the performance, three metrics are applied, i.e., Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and Spectral Angle Mapper (SAM). Among these metrics, the higher the values of PSNR and SSIM are, the better its performance is. The value of SAM is small, which means the spectral distortion is lower.

C. Implementation Details

For CAVE and Harvard dataset, the spectral response function of Nikon D700 camera[5] is utilized to synthesize HR RGB image Y by means of Eq. 1 [8], [10], [15], [27], [28]. Another dataset, Chikusei, we crop the training set into non-overlapping image with $200 \times 194 \times 128$. Since the number of training set is less, 64 patches are randomly cropped on each image. Each patch is augmented by randomly flip, rotation, and roll.

With respect to the parameters of model $\mathcal{C}$, the convolution kernels involved in the network are fixed as 3×3, except for after concatenation operation. The number of convolution kernels is defined as 64. We adopt ADAM optimizer with β_1 = 0.9 and β_2 = 0.99 to optimize the model $\mathcal{C}$. The initial learning rate is set to 10^{-4}. In the training stage, its values are decayed by half for every 30 epochs. For the parameters of model $\mathcal{F}$, we also employ ADAM optimizer to learn parameters. The learning rate is fixed as 5^{-5} in our study. The experiments are implemented on PyTorch framework with NVIDIA GeForce RTX 3090.

TABLE II
Ablation study for scale $\times 8$ on CAVE dataset.

Component	Different combinations of components			
ALBA module	×	√	×	√
Spectral context	×	×	√	√
PSNR	39.623	40.832	41.486	42.466
SSIM	0.9667	0.9713	0.9809	0.9855
SAM	6.267	5.667	4.981	4.412

TABLE III
Effect of different block size on the performance of the model.

Block size	2×2	3×3	4×4	5×5
PSNR	42.490	42.466	42.873	42.023
SSIM	0.9854	0.9855	0.9831	0.9789
SAM	4.423	4.412	4.415	4.780

D. Ablation Study

1) Block Size Analysis: The proposed symmetrical feature propagation model designs two key modules to improve performance, i.e., adaptive local block aggregation (ALBA) module and spectral context. To verify the effectiveness of these two modules, the model is tested for scale factor $\times 8$ on CAVE dataset by removing or adding modules. In the experiment, the blur kernel with size 13×13 is employed to downsample the HR HSI directly by convolution, and the Gaussian noise with mean 0 and variance 0.001 is added to attain the LR HSI. Table II describes the performance of the model when different modules are combined. Since the designed ALBA module and spectral context consider local and global information respectively, they are beneficial to feature representation and enhance the learning of similar content. As seen from the table, this combination produces better performance than any other module combinations. The experiment reveals the effectiveness and superiority of the proposed modules.

E. Study of Adaptive Local Block Aggregation Module

To investigate the effectiveness of each part of ALBA module, experiments are conducted for scale factor $\times 8$ on CAVE dataset from three aspects, i.e., block size, block number and

[1] http://www1.cs.columbia.edu/CAVE/databases/multispectral/
[2] http://www1.cs.columbia.edu/CAVE/databases/multispectral/
[3] http://naotoyokoya.com/Download.html
[4] https://www.l3harrisgeospatial.com/Data-Imagery/Satellite-Imagery/High-Resolution/WorldView-2
[5] http://www.maxmax.com/spectral response.htm

TABLE IV
PERFORMANCE COMPARISON OF SEVERAL METHODS UNDER BLUR KERNELS.

Blur kernel	Metric	LTTR	CMS	PZRes-Net	MHF-net	UAL	CoarseNet
K_1	PSNR	31.309	31.905	34.213	36.000	40.549	42.041
	SSIM	0.6884	0.7632	0.8461	0.8719	0.9698	0.9852
	SAM	31.228	25.335	22.183	12.488	10.789	4.461
K_2	PSNR	31.331	31.766	34.161	35.956	40.534	42.015
	SSIM	0.6893	0.7609	0.8477	0.8716	0.9698	0.9849
	SAM	31.220	25.311	21.952	12.501	10.847	4.485
K_3	PSNR	31.346	32.133	34.292	36.082	40.572	42.147
	SSIM	0.6891	0.7649	0.8468	0.8724	0.9702	0.9854
	SAM	31.150	25.150	22.038	12.446	10.751	4.437
K_4	PSNR	31.035	31.857	34.203	35.975	40.414	40.850
	SSIM	0.6838	0.7600	0.8455	0.8698	0.9694	0.9837
	SAM	31.117	25.262	21.989	12.567	10.945	4.731

TABLE V
EFFECT OF DIFFERENT BLOCK NUMBER ON THE PERFORMANCE OF THE MODEL.

Number of similar blocks	1	2	3
PSNR	41.796	42.381	42.466
SSIM	0.9803	0.9796	0.9855
SAM	4.795	4.707	4.412

TABLE VI
EFFECT OF DIFFERENT BLOCK FUSION COEFFICIENT ON THE PERFORMANCE OF THE MODEL.

Coefficient	1	Self-learning	Cosine similarity
PSNR	42.383	42.457	42.466
SSIM	0.9849	0.9807	0.9855
SAM	4.307	4.564	4.412

block fusion coefficient. In the experiment, we also adopt the same way as in Section IV-D to generate LR HSI.

When the block relation is established, the HR RGB image is spatially divided into various non-overlapping blocks with the same size. To this end, we set four blocks with different sizes to study the influence of performance. The numerical results are shown in Table III. One can observe that different block sizes have a greater impact on the performance of the model. In particular, when the block size is greater than 4, the performance changes greatly. The main reason is that too much dissimilar content is included in the larger blocks, which hinders aggregation learning. If the block size is fixed at small, more similar contents can be explored, which is well illustrated by the results in the table. Considering that too small block size takes more time to calculate the position and similarity, this paper therefore sets the block size to 3×3 to implement the following experiments.

1) Block Number Analysis: In separation-aggregation unit, some adjacent blocks participate in the establishment of block relation. During aggregating these blocks, to analyze how many highly similar blocks are selected to facilitate the learning of context content, this section sets three numbers of adjacent blocks for feature aggregation. Table V depicts the effect of different numbers of blocks on the performance of the model. It can be observed from the table that when the number of similar blocks involved in the establishment increases, the values of three metrics are significantly improved. It indicates that the more the number of participating blocks is, the more conducive to model learning is. However, too many blocks will cause parameters to be added during fusion. Under the compromise consideration, three adjacent blocks are picked to study.

2) Block Fusion Coefficient Analysis: When constructing the relation between the current block and the neighborhood blocks, the cosine similarity is used to calculate the similarity between them, and it is also set as the block fusion coefficient in the separation-aggregation unit to realize the block fusion. To examine the effectiveness of cosine similarity, we fix the block fusion coefficient to 1 or self-learning to observe the changes in performance. Table VI reports the results of different corresponding block fusion coefficients. Although the model can obtain ideal results when the fusion coefficient is self-learning or fixed to 1, it can not deal with the new input LR HSI image very well. In contrast, our method can dynamically regulate the block fusion coefficient using cosine similarity according to the change of the input image. It illustrates the superior of this coefficient.

F. Generalizability to Various Degradations

To analyze the generalization ability of the proposed method under diverse degradation conditions, experiments are performed on CAVE dataset by setting various blur kernels, noise levels, and downsampling ways.

1) Study of Blur Kernels: This section utilizes four blur kernels with size 13×13, 15×15, 17×17, and 21×21 to examine. Fig. 8 displays visual results of the corresponding blur kernel. For convenience of description, the four blur kernels are represented as K_1, K_2, K_3, and K_4 respectively. Similarly, we also adopt the same way as in Section IV-D to generate LR HSI through four blur kernels. The results of these competitors are displayed in Table IV. Since the predefined degenerate kernels in LTTR, CMS, PZRes-Net and MHF-net deviate from the real degenerate kernels, none of these methods can produce satisfactory SR results. Although UAL estimates the degradation kernel of the image, it can not well identify the difference information between the RGB image and the LR HSI. Importantly, this approach does not take

TABLE VII
PERFORMANCE COMPARISON OF SEVERAL METHODS UNDER NOISE LEVELS.

Variance	Metric	LTTR	CMS	PZRes-Net	MHF-net	UAL	CoarseNet
0.0005	PSNR	34.123	34.207	36.870	38.102	42.515	43.415
	SSIM	0.7889	0.8403	0.8947	0.9136	0.9800	0.9900
	SAM	26.672	21.129	18.760	10.136	8.727	3.857
0.001	PSNR	31.309	31.905	34.213	36.000	40.549	42.073
	SSIM	0.6884	0.7632	0.8461	0.8719	0.9698	0.9850
	SAM	31.228	25.335	22.183	12.488	10.789	4.491
0.002	PSNR	28.651	29.572	31.565	33.838	38.540	40.216
	SSIM	0.5674	0.6600	0.7820	0.8122	0.9535	0.9734
	SAM	36.052	29.932	25.704	15.502	13.436	5.485

TABLE VIII
PERFORMANCE COMPARISON OF SEVERAL METHODS UNDER DOWNSAMPLING APPROACHES.

Downsampling type	Metric	LTTR	CMS	PZRes-Net	MHF-net	UAL	CoarseNet
Convolution	PSNR	31.309	31.905	34.213	36.000	40.549	42.018
	SSIM	0.6884	0.7632	0.8461	0.8719	0.9698	0.9852
	SAM	31.228	25.335	22.183	12.488	10.789	4.445
Bicubic	PSNR	30.930	32.570	34.276	36.170	40.620	42.466
	SSIM	0.6840	0.7732	0.8514	0.8730	0.9700	0.9855
	SAM	31.261	24.888	21.607	12.474	10.799	4.412
Bilinear	PSNR	30.977	32.603	34.319	36.192	40.633	42.191
	SSIM	0.6856	0.7743	0.8517	0.8738	0.9703	0.9854
	SAM	31.202	24.838	21.612	12.439	10.772	4.375

full advantage of the inherent properties of the input image. Therefore, it produces relatively poor results. In contrast, the proposed method designs random blur kernels with various types and sizes. Besides, our method constructs a symmetrical structure, which effectively utilizes the content of two modalities. As a result, it clearly outperform all competitors.

2) Study of Noise Levels: To explore the performance of the comparison methods under different noise levels, this section appends Gaussian noise with variance of three values and mean 0 to the downsampled image. Table VII shows the comparison results of existing methods. Among these competitors, most methods assume clean image when LR HSIs are constructed, which makes them more sensitive to noisy images. As observed from this table, various noise levels have a great impact on the performance of these approaches. Although UAL alleviates this problem by adding post-processing, its performance is still low. Unlike these competitors, the proposed method sets random noise levels, which can well address real image with noise, thus achieving ideal results.

3) Study of Downsampling Approaches: Currently, there are usually two ways to downsample the HR image, including convolution and interpolation. This section applies common downsampling methods, namely, convolution, Bicubic interpolation and Bilinear interpolation, to investigate the performance of existing methods. Table VIII describes that the results for various downsampling strategies. As seen from this table, the performance of traditional methods is not stable. The models based on deep learning yield relatively stable results in comparison to them. However, some of these networks among them still fluctuate greatly in three metrics, which reveals that the robustness is poor. Our method takes into account various downsampling types. Thus, it surpasses all algorithms by a clear margin.

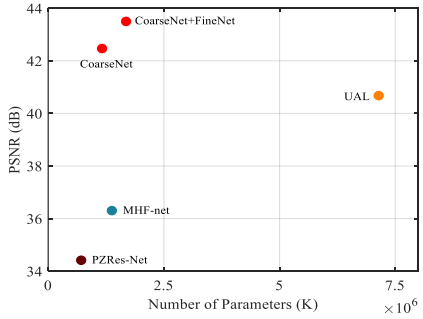

Fig. 9. PSNR performance versus number of parameter for scale factor ×32.

G. Generalizability to Different Datasets and Scales

In this section, the proposed model is compared with five methods on three datasets to verify the generalizability of the model. In the experiment, the blur kernel with size 13×13 is employed to blur the HR HSI, and the blurred image is downsampled by Bicubic interpolation. Then, we add Gaussian noise with mean 0 and variance 0.001 to generate the LR HSI.

Table IX provides the evaluation results of the mainstream models on three datasets. As for CAVE and Harvard dataset, it can be seen that the proposed CoarseNet is superior to other competitors in most cases, i.e., CoarseNet achieves remarkable performance in small scale factor and comparable results in large scale factor. After joining the FineNet, our model can effectively make up for this disadvantage. For instance, compared with the performance of UAL for scale factor ×8 on CAVE dataset, the values of CoarseNet in PSNR improve 1.8dB. Interestingly, its value is improved by at

TABLE IX
PERFORMANCE COMPARISON OF EXISTING APPROACHES ON VARIOUS DATASETS.

Dataset	Scale factor	Metric	LTTR	CMS	PZRes-Net	MHF-net	UAL	**CoarseNet**	**CoarseNet +FineNet**
CAVE	×8	PSNR	30.930	32.570	34.276	36.170	40.620	42.466	43.498
		SSIM	0.6840	0.7732	0.8514	0.8730	0.9700	0.9855	0.9863
		SAM	31.261	24.888	21.607	12.474	10.799	4.412	4.237
	×32	PSNR	36.954	37.967	34.193	35.112	39.891	38.050	40.134
		SSIM	0.9204	0.9528	0.8656	0.9208	0.9735	0.9732	0.9700
		SAM	16.806	11.145	21.700	8.701	10.385	5.278	5.240
Harvard	×8	PSNR	30.032	32.492	33.404	34.212	39.235	41.896	42.405
		SSIM	0.6994	0.7917	0.8558	0.8470	0.9602	0.9737	0.9748
		SAM	23.542	18.864	17.133	13.869	7.787	3.714	3.758
	×32	PSNR	34.818	38.774	33.086	36.495	39.818	39.648	40.684
		SSIM	0.8952	0.9500	0.8806	0.9345	0.9707	0.9670	0.9695
		SAM	14.850	6.520	16.991	6.482	6.088	4.135	4.245
Chikusei	×8	PSNR	–	–	34.650	32.363	36.982	38.005	–
		SSIM	–	–	0.8784	0.8363	0.8966	0.9551	–
		SAM	–	–	13.228	13.282	3.048	8.452	–
	×32	PSNR	–	–	36.428	26.476	34.005	35.785	–
		SSIM	–	–	0.8954	0.8441	0.8411	0.9403	–
		SAM	–	–	12.675	12.587	5.163	11.174	–

least 2.8dB, after adding FineNet to learn global information. Importantly, the proposed approach attains a better trade-off between performance and model size, which is shown in Fig. 9. Both numerical results and figures reveal that these competitors are slightly inferior to our method in terms of performance and parameters. Similarly, the proposed approach exhibits excellent performance on the Harvard dataset.

We also verify the models on remote sensing Chikusei dataset. Note that this dataset does not provide a corresponding spectral response function. Considering that LTTR and CMS need to utilize the spectral response function many times in the process of SR, the results of these methods are not shown. Besides, UAL designs an unsupervised adaptive network to optimize the model by minimizing the difference between them, including RGB image transformed by spectral response function and reference RGB image, downsampling HSI and input LR HSI. To ensure accurate results, we do not add FineNet during the test. Similar to the results of the other two datasets, it can be seen from this table that CoarseNet achieves desirable results on small scale factor without FineNet, while partial results obtain comparable performance on large scale factor.

Figs. 11 and 10 exhibit the spatial and spectral details of super-resolved HSI. To clearly present the results in space, the band of the super-resolved HSI is first selected, then the absolute difference between it and reference band is calculated. As seen from the figure, the proposed method obtains less superficial information in the magnified area, which means that the reconstructed image can generate clearer details. As for spectrum retention, we first randomly select a pixel in the spatial domain, and draw the spectral curve along the spectral dimension. According to the distribution of the curve, our method can be consistent with the reference curve in most cases. The experiments demonstrate that the proposed method has strong generalization ability in terms of cross datasets and different scales.

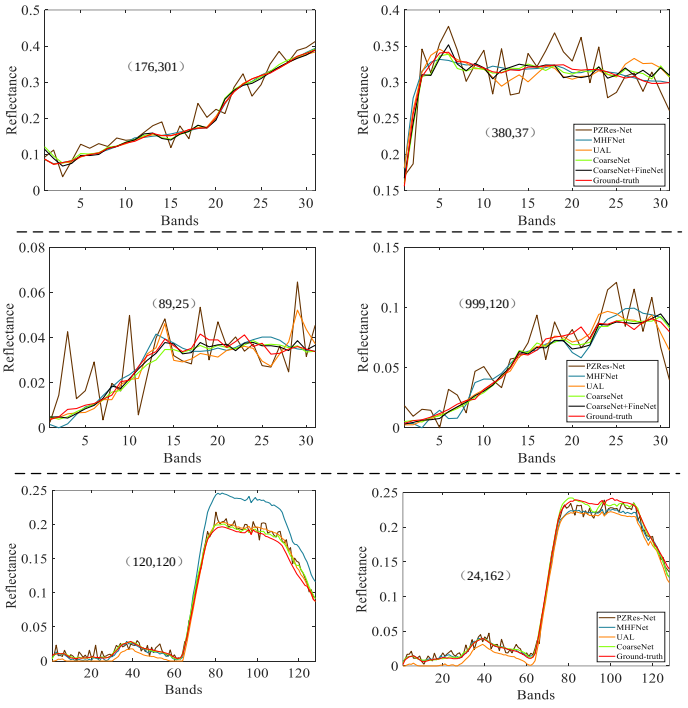

Fig. 10. Visual comparison of spectral distortion by selecting two pixels.

H. Application on Real Hyperspectral Image

In this section, we apply the proposed method to a real sample located at Roman Colosseum to demonstrate its applicability. Since the HR image for this dataset is not available, the algorithm proposed in [42] is employed to tackle this trouble. Concretely, we first randomly crop patch with size 36 ×36 × 8 from HSI, and acquire the corresponding RGB

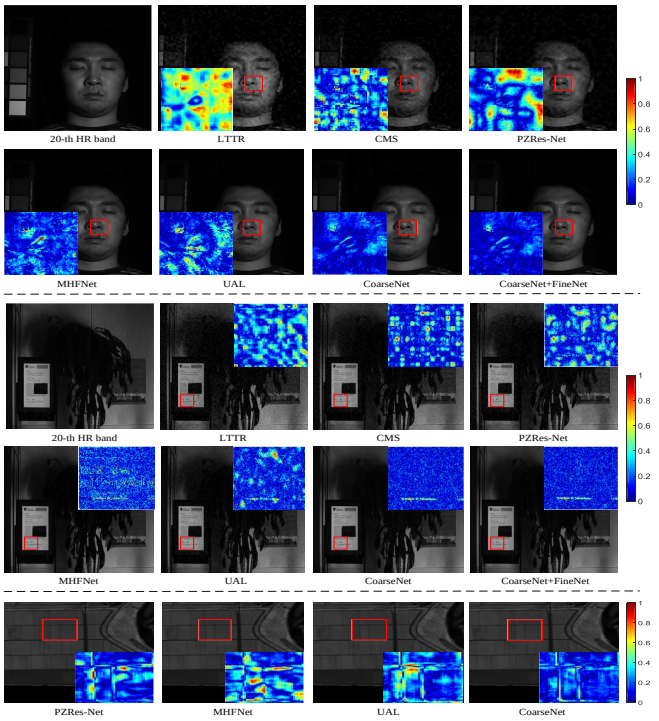

Fig. 11. Visual comparison in spatial domain for three images. The first to three lines represent the visual results of the 20-th band, 20-th band, and 80-th band, respectively.

Fig. 12. Visual comparison on real HSI dataset. We choose the 2-3-5 bands after SR to synthesize the pseudo-color image.

patch with size $144 \times 144 \times 3$. Then, the HSI patch and RGB patch are downsampled into the size $9 \times 9 \times 8$ and $36 \times 36 \times 3$, respectively. Finally, the downsampled patches and original patches are exploited as training pairs. To examine the applicability, three existing models are adopted. Since there is no spectral response function, we take above similar approach and show only partial results of the methods. Fig. 12 exhibits the super-resolved results for existing methods. We find that some images appear smoothing or ringing results. However, the proposed CoarseNet recovers clear textures. It reveals our work can effectively tackle images in real scenes and has certain practicability.

V. Conclusion

In this paper, we develop a coarse-to-fine learning approach for HSI SR. In coarse stage, the method promotes the information propagation and fusion between LR HSI and RGB image through multi-step propagation strategy. This structure forms a symmetric pattern, which encourages the two forms to better retain their particularity. Meanwhile, an adaptive local block aggregation module is designed to adaptively aggregate blocks with high texture similarity, so as to achieve feature enhancement. On this basis, we introduce a back projection refinement network in fine stage to learn image-specific details globally. The experimental results show that the proposed approach is superior to the existing methods in quantitative and qualitative evaluation. As the proposed method only models the similar content in spatial domain, it ignores the relation between spectra. In the future, we will explore this similarity in spatial domain and spectral domain, further realizing feature enhancement.

References

[1] Y. Yuan, L. Dong, and X. Li, "Hyperspectral unmixing using nonlocal similarity-regularized low-rank tensor factorization," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2021.

[2] Y. Yuan, C. Wang, and Z. Jiang, "Proxy-based deep learning framework for spectral–spatial hyperspectral image classification: Efficient and robust," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–15, 2021.

[3] S. Valero, P. Salembier, and J. Chanussot, "Object recognition in hyperspectral images using binary partition tree representation," *Pattern Recognit. Lett.*, vol. 56, pp. 45–51, 2015.

[4] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral face recognition with spatiospectral information fusion and PLS regression," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1127–1137, 2015.

[5] G. Tochon, J. Chanussot, M. Dalla Mura, and A. L. Bertozzi, "Object tracking by hierarchical decomposition of hyperspectral video sequences: Application to chemical gas plume tracking," *IEEE Trans. Geosci. Remote Sensing*, vol. 55, no. 8, pp. 4567–4585, 2017.

[6] S. Mei, Y. Geng, J. Hou, and Q. Du, "Learning hyperspectral images from RGB images via a coarse-to-fine CNN," *Sci. China-Inf. Sci.*, vol. 65, no. 5, pp. 1–14, 2022.

[7] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3D full convolutional neural network," *Remote Sensing*, vol. 9, no. 11, pp. 1139, 2017.

[8] L. Zhang, W. Wei, C. Bai, Y. Gao, and Y. Zhang, "Exploiting clustering manifold structure for hyperspectral imagery super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5969–5982, 2018.

[9] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, 2019.

[10] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, 2019.

[11] j. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "A truncated matrix decomposition for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 8028–8042, 2020.

[12] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, 2018.

[13] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Trans. Circuits Syst. Video Technol.*, 2021.

[14] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, Y. Shi, and J. Chanussot, "Edge-conditioned feature transform network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, 2021.

[15] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 3073–3082.

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

[17] J. Xue, Y.-Q. Zhao, Y. Bu, W. Liao, J. C. Chan, and W. Philips, "Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 3084–3097, 2021.

[18] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3631–3640.
[19] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, 2022.
[20] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projectinetworks for single image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4323–4337, 2021.
[21] Z. Wang, J. Chen, and S. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, 2021.
[22] T. Isobe, X. Jia, S. Gu, S. Li, S. Wang, and Q. Tian, "Video super-resolution with recurrent structure-detail network," in *European Conf. on Comput. Vision*, 2020, pp. 645–660.
[23] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse dirichlet-net for hyperspectral image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2511–2520.
[24] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "Fusion-net: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 7565–7577, 2020.
[25] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 87–102.
[26] B. Pan, Q. Qu, X. Xu, and Z. Shi, "Structure-color preserving network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–12, 2022.
[27] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Trans. Image Process.*, vol. 30, pp. 1423–1438, 2021.
[28] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by ms/hs fusion net," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1585–1594.
[29] X.-H. Han, Y. Zheng, and Y. W. Chen, "Multi-level and multi-scale spatial and spectral fusion CNN for hyperspectral image super-resolution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2019, pp. 4330–4339.
[30] Z. Zhe, Y. Zheng, and X.-H. Han, "Unsupervised multispectral and hyperspectral image fusion with deep spatial and spectral priors," in *Proceedings of the Asian Conference on Computer Vision Workshops*, 2020, pp. 31–45.
[31] X.-H. Han, B. Shi, and Y. Zheng, "SSF-CNN: Spatial and spectral fusion with cnn for hyperspectral image super-resolution," in *Proc. Int. Conf. Image Process.*, 2018, pp. 2506–2510.
[32] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 7, pp. 5953–5965, 2021.
[33] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.
[34] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[35] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, "Deep blind hyperspectral image fusion," in *Proc. IEEE Int. Conf. Comput Vision*, 2019, pp. 4150–4159.
[36] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 5754–5768, 2021.
[37] Q. Wang, Q. Li, and X. Li, "A fast neighborhood grouping method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 6, pp. 5028–5039, 2020.
[38] Q. Wang, Q. Li, and X. Li, "Hyperspectral image super-resolution using spectrum and feature context," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 11276–11285, 2021.
[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc IEEE Int. Conf. Comput Vision*, 2017, pp. 2242–2251.
[40] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, 2010.
[41] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 193–200.
[42] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, 2018.

Qiang Li received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China in 2022.

He is currently a postdoc with the School of Electronic Engineering, Xidian University, Xi'an. His research interests include remote sensing image processing and computer vision.

Maoguo Gong (M'07-SM'14) received the B.S. and Ph.D. degrees in electronic science and technology from Xidian University, Xi'an, China, in 2003 and 2009, respectively.

Since 2006, he has been a Teacher with Xidian University. In 2008 and 2010, he was promoted as an Associate Professor and as a Full Professor, respectively, both with exceptional admission. His current research interests include computational intelligence with applications to optimization, learning, data mining, and image understanding.

Prof. Gong was a recipient of the Prestigious National Program for the support of Top-Notch Young Professionals from the Central Organization Department of China, the Excellent Young Scientist Foundation from the National Natural Science Foundation of China, and the New Century Excellent Talent in University from the Ministry of Education of China. He is the Vice Chair of the IEEE Computational Intelligence Society Task Force on Memetic Computing, an Executive Committee Member of the Chinese Association for Artificial Intelligence, and a Senior Member of the Chinese Computer Federation. He is also an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.