RGB-Induced Feature Modulation Network for Hyperspectral Image Super-Resolution

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

***Abstract*— Super-resolution (SR) is one of the powerful techniques to improve image quality for low-resolution (LR) hyperspectral image (HSI) with insufficient detail and noise. Traditional methods typically perform simple cascade or addition during the fusion of the auxiliary high-resolution (HR) RGB and LR HSI. As a result, the abundant HR RGB details are not utilized as a priori information to enhance the HSI feature representation, leaving room for further improvements. To address this issue, we propose an RGB-induced feature modulation network for HSI SR (IFMSR). Considering that similar patterns are common in images, a multi-corresponding patch aggregation is designed to globally assemble this contextual information, which is beneficial for feature learning. Besides, to adequately exploit plentiful HR RGB details, an RGB-induced detail enhancement (RDE) module and a deep cross-modality feature modulation (CFM) module are proposed to transfer the supplementary materials from RGB to HSI. These modules can provide a more direct and instructive representation, leading to further edge recovery. Experiments on several datasets demonstrate that our approach achieves comparable performance under more realistic degradation condition. Our code is publicly available at https://github.com/qianngli/IFMSR.**

***Index Terms*— Detail enhancement, feature modulation, hyperspectral image (HSI), super-resolution (SR).**

I. Introduction

THE spatial and material attributes of objects can be quantitatively analyzed by using hyperspectral images (HSIs), which immensely advances the performance of classification [1] and object detection [2]. However, the image quality is degraded in the process of obtaining HSI due to the influence of sensor system and external conditions [3], [4]. As a result, the degraded image evidently loses high-frequency details of the real scene, which is not conducive to the effective discrimination of the objects. It critically restricts the ability of accurate perception for HSI. Therefore, there is an urgent to develop corresponding techniques to enhance the quality of low-resolution (LR) image, which inevitably reinforces the performance of subsequent interpretation and analysis.

HSI super-resolution (SR), as an economical and effective technique to improve image quality, has attracted the attention of numerous researchers in recent years [5], [6], [7], [8], [9]. Since spectral imagers commonly generate RGB image and HSI simultaneously, considering the extreme spatial resolution of RGB image, there are various algorithms that combine LR HSI and high-resolution (HR) RGB image to perform HSI SR [9], [10], including traditional methods [11], [12], [13], [14] and deep learning-based methods [15], [16], [17], [18]. The performance of related algorithms is continuously improving. Existing traditional approaches construct features manually and use different priori knowledge to design the model. Since the feature representation ability by handcrafted is limited, these methods are not able to recover the more details for realistic LR images, which makes the traditional methods less robust.

Benefiting from the powerful representation of convolutional neural networks [19], [20], researchers have mainly exploited deep learning to build models for HSI SR. Compared to traditional methods, deep learning-based methods have made remarkable progress in dealing with real LR images. For instance, some existing methods [21], [22], [23] aim at cascading RGB image with LR HSI to model. Actually, RGB image and LR HSI have high spatial resolution and rich spectral information, respectively. The remarkable property can provide a more direct and instructive representation, which is favorable for edge recovery. Nonetheless, this type of approach ignores the exploration of this aspect, resulting in miserable performance. To solve this problem, the researchers input RGB image and LR HSI into the networks in parallel [24], [25], [26], and two branches are designed to learn the features of each modality separately, and the fusion operation is conducted at the back end. Similar works are also found in other fields such as [27] and [28]. Although these networks sufficiently mine the internal information of their respective modality, they do not skillfully borrow the RGB image with full appearances to transfer the detailed contents to HSI modality for feature learning. Therefore, how to leverage the intrinsic advantages of RGB image to assist the HSI branch to capture more discriminative features requires further research.

Manuscript received 6 December 2022; revised 30 March 2023 and 23 April 2023; accepted 15 May 2023. Date of publication 18 May 2023; date of current version 1 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U21B2041 and Grant 61825603 and in part by the National Key Research and Development Program of China under Grant 2020YFB2103902. *(Corresponding author: Qi Wang.)*

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

Maoguo Gong is with the State Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, Xi'an 710071, China (e-mail: gong@ieee.org).

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

Digital Object Identifier 10.1109/TGRS.2023.3277486

SR is a low-level task that requires more contextual information to get clearer contents. To this end, some researchers design nonlocal information fusion of multiple similar patches to achieve better performance [29], [30], [31]. For instance, Li et al. [30] propose a temporal multi-correspondence aggregation module. It aggregates the patch of the current frame and the patches with greater similarity in the adjacent frame. Considering that the fusion of similar patches with the same LR cannot make the network directly perceive the image texture with higher resolution, Zhou et al. [31] design a cross-scale patch aggregation module to achieve cross-size information fusion. Inspired by these modules, we also adopt similar strategies to find out several patches from the input RGB image that has high spatial resolution. In feature extraction stage, the similar patches are aggregated on the two branches in a shared way to enhance feature representations.

Motivated by these discoveries, this article proposes an RGB-induced feature modulation network for HSI SR (IFMSR). Specifically, we design a dual-branch model to form a symmetrical structure. First, we directly divide raw RGB image into identical patches to compute relational index list in input stage instead of feature extraction stage, which is helpful to generate more accurate list. According to the constructed list, this knowledge is shared at the feature level, and the nonlocal contents of the respective modality are explored on the two branches in same resolution. Considering that abundant HR RGB details can provide a more direct and instructive representation, an RGB-induced detail enhancement (RDE) module is proposed. It employs kernel generation in RGB modality to obtain region-aware dynamic filters for image-specific. Then, the region-aware dynamic filters are utilized to guide encoding in HSI modality with different depths. Moreover, the cues generated by the multimodal hybrid features are used for cross-modal guidance to ensure their specificity. To produce super-resolved HSI, the two modalities must be into a unified form during deep feature extraction. In this process, the RGB features are the auxiliary signals relative to the HSI features. Therefore, we introduce a deep cross-modality feature modulation (DCFM) module, which integrates two modality information effectively through feature transformation. Different from previous works [32], [33] and concatenation operation, this module uses the content of RGB branch to guide the fusion learning of HSI branch. It is actually a guided form of image fusion. Experimental results reveal that our approach can yield comparable performance in both quantitative and qualitative aspects. In summary, the contributions of this article are threefold.

1) We propose an RGB-guided feature modulation network for HSI SR. In particular, we provide a different perspective to model the dependence of two modalities in feature representations.
2) We design an RDE module and a deep cross-modality feature modulation (DCFM) module to transfer the detail supplementary information from RGB modality to HSI modality with different depths, so as to increase the detailed exploration of HSI features.
3) We evaluate the effectiveness of the proposed IFMSR, including RDE and DCFM module. Experiments demonstrate that our IFMSR can better handle LR images against existing methods on both synthetic and real datasets.

The rest of this article is organized as follows: Section II reviews traditional and deep learning-based methods for HSI SR. Section III provides the details of the proposed method. Section IV analyzes and discusses the experiments. Finally, Section V gives the conclusion of this work.

II. Related Work

SR is one of the important techniques to improve image quality for LR HSI. Currently, the existing approaches are roughly divided into two types, i.e., traditional and deep learning-based methods.

A. Traditional Methods

Under the assumption of sparse, low-rank, etc., a large number of traditional methods based on various shallow priori knowledge have been proposed. Wei et al. [11] utilize several spectral atoms to represent each spectrum, and the dictionary is learned from HR RGB image to promote as much spatial consistency as possible with LR HSI. Given the spectral response, Wycoff et al. [34] transform the SR problem into a non-negative sparse decomposition and adopt an alternating multiplier to optimize model. Dong et al. [13] employ a non-negative dictionary to study the non-negative spectral cardinality. Akhtar et al. [12] first learn non-negative dictionary and then introduce greedy algorithm to estimate the coefficient of local blocks. These above methods based on matrix factorization convert 3-D data into 2-D data. Although the information presented is the same in both ways, it destroys the intrinsic nature of HSI. Obviously, it cannot handle spectral material to improve feature learning in spatial content.

Since HSI can be naturally represented by 3-D tensor, several methods via tensor decomposition have been developed. For instance, Dian et al. [35] first time to utilize nonlocal sparse tensor decomposition to solve HSI SR. Zhang et al. [36] propose a low-rank Tucker decomposition for graph regularization to retain spatial consistency and spectral smoothness. Xu et al. [37] establish a model using a higher order coupled tensor with graph-Laplacian regularization. These traditional algorithms are constructed by handcrafted features. In the challenging scenarios with strong noise, these approaches exhibit poor generalization due to the limited ability of shallow feature representations.

B. Deep Learning-Based Methods

Inspired by the great success of deep learning, many approaches have been proposed for HSI SR. As for supervised approaches, most previous works refer to natural image SR algorithms to build frameworks based on HSI properties. For example, Han et al. [21] cascade HR RGB image and upsampled LR HSI together to jointly explore spectral and spatial contents. Zhang et al. [22] exploit similar scheme to design the network, and the super-resolved image is estimated by spatial and spectral reconstruction network. A slight difference

is that Hu et al. [17] downsample RGB image and integrate them with LR HSI into the model as a whole. Besides, the spectral and spatial preservation modules force the model to generate reconstructed image. These methods mentioned above aggregate the two modalities together during input, which prematurely destroys the priori properties of the images, i.e., high spatial resolution for RGB image and rich spectral information for LR HSI.

To effectively exploit these properties, the parallel input manner is considered. For instance, Han et al. [38] propose a multiscale spatial and spectral DCNN to investigate a HR spatial structure reservation and spectral reservation in a parallel way. Xie et al. [39] integrate the observation model and image priori learning into single structure at the end. Later, Zhang et al. [16] construct a general priori model to capture spatial-spectral information in coarse stage, and propose an adaptation module to depict image-specific details in fine stage. Dong et al. [40] applies both domain knowledge likelihood and deep image prior, and develop a model-guided deep network. These approaches mainly fuse the spectral information extracted from HSI and spatial information extracted from RGB image by means of cascade or addition. Intuitively, rich RGB image details are not applied as a priori information to enhance the feature representation for HSI, leaving room for further improvements.

III. Proposed Method

Given LR HSI $X \in \mathbb{R}^{B\times w\times h}$ and the corresponding RGB image $Z \in \mathbb{R}^{3\times W\times H}$, SR task aims to obtain super-resolved image $Y \in \mathbb{R}^{B\times W\times H}$ according to upscale factor r. Here, B is the number of total bands in the image. W and H are the width and height for images Y and Z. w and h are the width and height for image X. Since adjacent bands present high correlation, Wang et al. [41] only utilize several bands to achieve HSI SR, which has been proven to attain better performance and reduce the memory footprint. Motivated by novel input pattern, in our article, only two adjacent bands X_{k-1}, X_{k+1} and current band X_k are selected and combined to restore single band, where $k \in \{1, \ldots, B\}$ indicates band index. Then, the recurrent strategy is applied to obtain each super-resolved band Y_k. The relationship among these images are follows:

$$Z = TX \tag{1}$$

$$X_k^{\text{cat}} = \left[X_{k-1}, X_k, X_{k+1}\right] \tag{2}$$

$$Y_k = \mathcal{F}\left(X_k^{\text{cat}}, Z, r\right) \tag{3}$$

where T is the spectral response function that is the inherent system parameter of the sensor, and $\mathcal{F}(.)$ is designed model function.

A. Overview

Since multiple bands are divided in a certain spectral range, HSI has high spectral resolution but considerably low spatial resolution [3]. In contrast, RGB image shows clear textures and edges. Therefore, the researchers adopt auxiliary RGB image to improve the performance for HSI SR. Despite previous approaches fuse the spectral information extracted from HSI and spatial information extracted from RGB image, these models do not full use the RGB image with rich appearances to transmit the detailed information to HSI modality for feature learning. In other words, the features from RGB image are not effectively utilized to induce the model to modulate the HSI branch.

To be able to take advantage of this salient priori information, in our article, we propose an IFMSR, which is shown in Fig. 1. Concretely, two adjacent bands X_{k-1}, X_{k+1} relative to the current band X_k (i.e., spectral context) are selected to reconstruct band using RGB image. By doing so, it can explicitly reduce the memory footprint. Meanwhile, the spectral context can be used to promote information complementarity. In terms of data input, two modalities are fed into the model in parallel, and the intrinsic knowledge of the respective modality is explored through two branches. In addition to using spectral context, i.e., adjacent bands, we also design a multi-correspondence patch aggregation (MCPA) module to increase the learning of similar patterns in global perspective, so as to achieve spatial context aggregation. During feature extraction, an RDE module is proposed, which serves as a bridge to transmit the abundant information for RGB image to HSI modality with different depths. Moreover, the multimodal hybrid features are attached to the two branches separately to ensure their specificity. After obtaining deep features through two branches, we aggregate the multiple hierarchical features together, and introduce the deep cross-modality feature modulation (DCFM) module to further boost the deep feature representations in HSI branch. Finally, the super-resolved HSI is obtained after one convolution layer.

B. Multi-Correspondence Patch Aggregation

When the spectral imager generates HSI, it commonly obtains the corresponding RGB image. How to address RGB image with abundant color and texture to guide HSI SR is extremely challenging dilemma. To fully utilize raw RGB image, we deal with this image to generate more meaningful materials. Generally, some patches in the whole image appear similar patterns. These patches can benefit detail restoration, which has been verified in several previous low-level tasks [30], [42]. Although Li et al. [10] consider contextual information in image, it only focuses on local content, and ignores other similar patterns in global perspective. To achieve this end, a MCPA module is developed, as shown in Fig. 2.

To easily conduct this procedure, the image Z is divided into identical patches P with non-overlapping size $S \times S$ in space. Note that the dimension of the input image is padded if equal size blocks cannot be obtained. Let the position of current patch P_i be (x, y), where i is patch index. We compute the correlation between current patch P_i and other patches $P_j (i \neq j)$ by traversal way in global perspective, and sort these values in descending order. Three patches with corresponding large correlation $C_{i,1}$, $C_{i,2}$, and $C_{i,3}$ are selected as candidate patches, i.e.,

$$C_i(j) = \frac{\left\langle P_i, P_j \right\rangle}{\|P_i\|_2 \times \left\|P_j\right\|_2} \tag{4}$$

$$\left[C_{i,1}, C_{i,2}, C_{i,3}\right] = \text{sort}(C_i, 3). \tag{5}$$

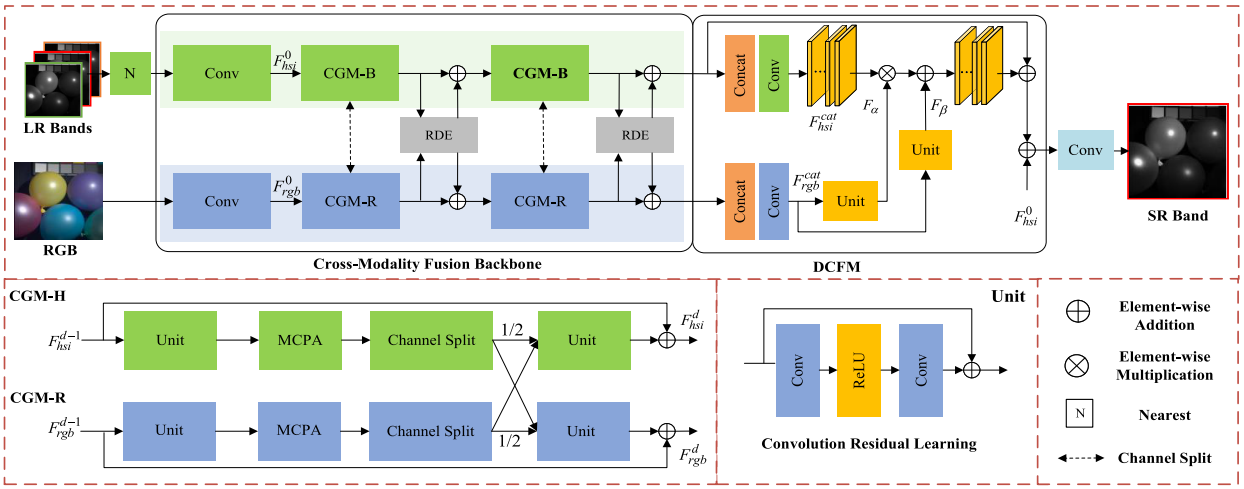

Fig. 1. Proposed the architecture of the proposed RGB-induced feature modulation network for HSI SR (IFMSR).

Accordingly, the positions of three patches are obtained, leading to relational index list, i.e., $(C_{i,1}, Z_{i,1})$, $(C_{i,2}, Z_{i,2})$, and $(C_{i,3}, Z_{i,3})$. Compared with dealing with intermediate features directly, more accurate relational index lists can be produced by analyzing the raw RGB image. After generating index list for each patch, we need to assemble these patches during feature extraction. Concretely, we upsample LR HSI to the same size as RGB image to explicitly exploit these patches with high correlation. During feature extraction in each modality, similarly, the intermediate features are divided into identical patches with non-overlapping size. Based on the obtained relational index list, each patch and its three highly correlated patches are aggregated for same manner, forming new features H_i' and $R_i' \in \mathbb{R}^{N\times S\times S}$ for HSI and RGB modality, i.e.,

$$R_i' = \mathcal{M}\left(\left[R_i, R_{i,l} * W_{r,l}\right]\right) \tag{6}$$

$$H_i' = \mathcal{M}\left(\left[H_i, H_{i,l} * W_{h,l}\right]\right) \tag{7}$$

where $l = 1, 2, 3$. N is the number of feature maps. W is learnable weight. R_i and $R_{i,l}$ represent feature maps for current patch and its three highly correlated patches in the RGB modality. H_i and $H_{i,l}$ represent feature maps for current patch and its three highly correlated patches in the HSI modality. $\mathcal{M}(\cdot)$ is convolution operation with kernel size 3×3. By doing so, it can enhance the feature representation ability of the each modality in global perspective.

C. RDE Module

Among the low-level encoder features, RGB features contain more detailed information (i.e., texture and color), which can provide a more direct and instructive representation than HSI features. This facilitates the feature learning during encoding. Nevertheless, previous works only simply fuses the features of each modality, and do not exploit this remarkable behavior to induce the model that encourages the feature exploration of HSI modality. To tackle this issue, an RDE module is proposed to generate region-aware dynamic filter to guide encoding in HSI modality. The structure of RDE module is illustrated in Fig. 3.

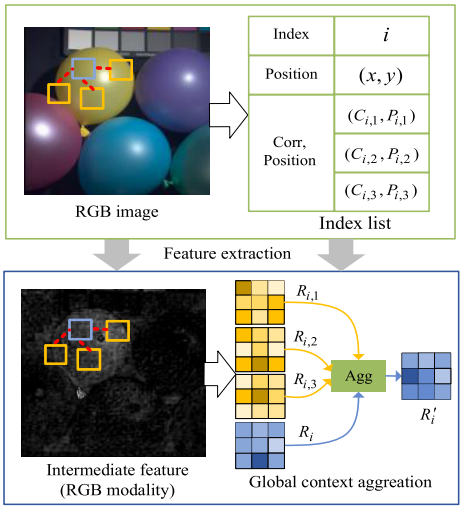

Fig. 2. Illustration of MCPA module. For clear description, the processing of patch aggregation only in RGB modality is shown.

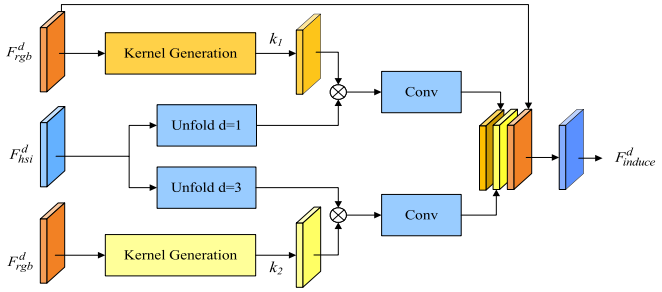

Fig. 3. The architecture of RGB-induced detail enhancement (RDE) module.

Object sometimes takes up a larger size in an image. If a small convolution kernel is adopted to extract the features, it cannot make use of these adjacent contents with similar patterns. Considering this trouble, dilated convolution is introduced to increase the receptive field in the HSI modality. To induce feature learning in HSI modality, we employ kernel generation in RGB modality to obtain region-aware dynamic

filters for image-specific. Suppose that $F^d_{\text{rgb}} \in \mathbb{R}^{N\times W\times H}$ and $F^d_{\text{hsi}} \in \mathbb{R}^{N\times W\times H}$ represent the input features of RGB and HSI modality in the dth RDE module, respectively. The dilated convolution captures the affluent data at multiple scales in HSI modality with the kernel generation to force model modulation, i.e.,

$$k_{\text{rate}} = \mathcal{N}\left(\mathcal{M}\left(F^d_{\text{rgb}}\right)\right) \tag{8}$$

$$F^d_{\text{induce}} = \mathcal{M}\left(\left[F^d_{\text{rgb}}, \mathcal{M}\left(\mathcal{F}_{\text{unfold}}\left(F^d_{\text{rgb}}, \text{rate}\right) \circledast k_{\text{rate}}\right)\right]\right) \tag{9}$$

where F^d_{induce} denotes induced features ($d = 1, 2$). $\circledast$ is an adaptive convolution operation. rate represents dilation rate for dilated convolution (rate $= 1, 3$). $\mathcal{F}_{\text{unfold}}(\cdot)$ extracts the features for local patch by sliding. $\mathcal{N}(\cdot)$ indicates convolution operation with kernel size 1×1. This module takes the features contained in the RGB branch as a priori information and transmits the detailed supplementary. It can induce the feature representations of HSI branch in the encoding, and suppress the incompatibility between the modalities.

D. Deep Cross-Modality Feature Modulation Module

To produce super-resolved HSI, the two modality must be into a unified form during deep feature extraction. In this process, the RGB features are the auxiliary signals relative to the HSI features. For this reason, we propose a DCFM module, as illustrated in Fig. 1. This module modulates the RGB features and obtains the affine transformation coefficient to further enforce deep features in HSI modality.

Given the deep RGB features $F^{\text{cat}}_{\text{hsi}} \in \mathbb{R}^{N\times W\times H}$ and HSI features $F^{\text{cat}}_{\text{rgb}} \in \mathbb{R}^{N\times W\times H}$ after multiple hierarchical feature concatenation, the DCFM module learns a mapping function $\mathcal{Q}$ to yield affine transformation parameters set $(F_\alpha, F_\beta) \in \mathbb{R}^{N\times W\times H}$, which is defined as

$$\left(F_\alpha, F_\beta\right) = \mathcal{Q}\left(F^{\text{cat}}_{\text{rgb}}\right). \tag{10}$$

With the estimated affine transformation parameters (F_α, F_β), the corresponding HSI features $F^{\text{cat}}_{\text{hsi}}$ are modulated by

$$F_{\text{rh}} = F^{\text{cat}}_{\text{hsi}} \otimes F_\alpha \oplus F_\beta \tag{11}$$

where $\oplus$ represents element-wise addition. Finally, the information both two modalities is effectively integrated through feature transformation.

E. Model Training

Existing works normally adopt Gaussian to degrade HR image during constructing label pairs. In real-world scenarios, image degradation is complicated by noise, blur, and other factors. The performance of traditional SR models using only a single type of kernel is significantly degraded for realistic degraded images. SR under unknown degradation is more challenging than traditional SR under simple degradation. Inspired by Li et al. [10] and Zhang et al. [43], anisotropic and isotropic Gaussian are employed to randomly generate different kernels with five sizes. Here, the range of rotation angle is set to $[0, \pi]$ for anisotropic Gaussian kernel, and the range of kernel width is fixed at $[0.2, r]$. As for isotropic Gaussian kernel, the range of kernel width is set to $[0.2, r]$. Then, HR HSI X' is downsampled by directly convolution, or is convoluted and interpolated. Gaussian noises with various levels are attached to the downsampled image. Finally, the label pair $\{X, X'\}$ is obtained. The training data $\mathcal{D}$ is constructed using the above steps. The model is optimized by loss function $\mathcal{L}$ in terms of synthetic data, i.e.,

$$\mathcal{L}(k) = \min\|X_k - Y_k\|_1 \tag{12}$$

where the subscript represents the band index in image, and $\|\cdot\|_1$ is $L1$ nrom.

IV. Experiments

To evaluate the effectiveness of the proposed method, we provide sufficient experiments in several aspects. First, the dataset and implementation details are presented. Then, the key modules are analyzed and discussed to verify the effectiveness. Finally, the proposed approach is compared with existing works under different conditions.

A. Datasets

1) CAVE: The CAVE dataset[1] was captured by Apogee Alta U260 from 400 to 700 nm at 10 nm steps [44]. The dataset contains 32 scenes, which are divided into five sections, including stuff, skin and hair, paints, etc. Each scene has 31 bands, where the resolution is 512×512 pixels. In addition, the corresponding RGB images are provided in the dataset. In our article, 80% images are randomly selected as training set and the rest as test set.

2) Harvard: The Harvard dataset[2] was collected in indoor and outdoor scenes under daylight illumination by Nuance FX, CRI Inc., from 420 to 720 nm at 10 nm steps [45]. Compared with CAVE dataset, all the images are of static scenes, where each scene consists of 31 bands with 1392×1040 pixels. Similarly, we adopt the above way to divide this dataset.

3) Chikusei: The Chikusei dataset[3] was taken by Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas from 363 to 1018 nm [46]. Unlike the above dataset, there is only one scenario in this dataset. The scene involves 2517×2335 pixels and the ground sampling distance is 2.5 m. We crop the top left of the HSI ($2000 \times 2335 \times 128$) as the training set, and other content as the test set. To obtain extra samples, the training set is divided into non-overlapping images with size $200 \times 194 \times 128$.

4) Sample of Roman Colosseum: The dataset[4] was obtained by WorldView-2 over Roman Colosseum area. Similarly, the dataset is a real-world scenario. The scene contains eight bands with 419×658 pixels, and the resolution of corresponding RGB image is 1676×2632 pixels. To evaluate the performance in this dataset, the top left of the HSI ($209 \times 658 \times 8$) and HR RGB image ($836 \times 2632 \times 3$) are cropped to train model, and other part is selected to test.

[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

TABLE I
COMPARISON PERFORMANCE BETWEEN TWO PATCH AGGREGATION MANNERS

Patch aggregation	PSNR	SSIM	SAM
Local manner in [10]	43.68	0.922	3.70
Global manner in our paper	43.83	0.992	3.69

TABLE II
PERFORMANCE OF DIFFERENT DEEP FEATURE EXTRACTION MODULES

Type	PSNR	SSIM	SAM
Single concatenation	43.30	0.991	3.87
Hierarchical concatenation	43.31	0.991	3.89
DCFM	43.83	0.992	3.69

TABLE III
ABLATION STUDY BY SETTING DIFFERENT COMBINATIONS OF MODULES

Module	Different combinations of modules					
DCFM	✔	✗	✔	✔	✗	✔
RDE	✗	✔	✔	✗	✔	✔
MCPA	✗	✗	✗	✔	✔	✔
PSNR	40.67	41.91	42.43	42.37	43.14	43.48
SSIM	0.985	0.988	0.989	0.990	0.991	0.992
SAM	5.12	4.55	4.31	4.17	3.91	3.75
Parameter	761K	970K	1135K	1351K	1561K	1725K

B. Implementation Details

Since the spectral response function is known on the CAVE and Harvard datasets, the corresponding RGB images are generated using it. Other datasets with unknown spectral response function, Chikusei and Sample of Roman Collosseum, obtain corresponding RGB images via the position of pixels. Then, we random crop 64 patches with the size $12r \times 12r$ from the each image in the training set, where r is upscale factor. All patches are augmented by random flip, rotation, and roll. Then, these patches are downsampled by the above strategy in Section III-E to yield LR HSIs. In the test stage, we adopt anisotropic Gaussian to generate kernel, so as to blur the HR HSI images. Each kernel is determined by a covariance matrix α, which is defined as

$$\alpha = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \tag{13}$$

where the ranges of these parameters are $\lambda_1 \in U[2, 5]$, $\lambda_2 \in U[8, 12]$, and $\theta \in U[0, \pi]$, respectively. Then, the blur images are downsampled and added Gaussian noise with mean 0 and variance 0.001. Finally, the test set is obtained.

With respect to the parameters of the network $\mathcal{F}$, the convolution kernels involved in the model are set to 3×3, and its number is fixed at 64, except for after concatenation. The model is optimized by ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 10^{-4} in our study, and its value decays by half for every 30 epoch. Experiments are performed on the PyTorch framework with NVIDIA GeForce GTX 1080 GPU. To evaluate the performance, peak signal-to-noise ratio (PSNR), structural similarity (SSIM), spectral angle mapper (SAM), relative dimensionless global error in synthesis (ERGAS), and root mean-squared error (RMSE) are exploited. Among these metrics, the higher the PSNR and SSIM values, the better the performance. The lower the SAM, ERGAS, and RMSE values, the better the reconstruction quality.

C. Study of Multi-Correspondence Patch Aggregation

Raw HR RGB image provides rich colors and textures. Therefore, we process this image to generate more meaningful materials. Concretely, patches with similar patterns are assembled in a global perspective to achieve feature enhancement. To compare the effectiveness of the proposed MCPA module, we introduce the way in a local perspective in [10] to aggregate adjacent contents. Table I shows the comparison performance between two aggregation manners. The approach [10] only calculates the current patch and its three adjacent patches. It ignores the remaining patterns in farther part. In contrast, our module searches globally for similar patterns. Moreover, since the two modalities have extremely similar contents, the designed module is applied to the HSI branch simultaneously. Experimental results show that the enhanced module in our article has a better advantage.

D. Study of Deep Cross-Modality Feature Modulation

In the deep feature extraction, the two modalities are assembled into a unified form to better obtain the final super-resolved HSI. In this process, we propose a DCFM module in bottleneck stage. To examine the performance improvement of DCFM module on the model, we employ two typical modules in this bottleneck stage to replace it, i.e., *single concatenation* and *hierarchical concatenation*. Here, the last features from two modalities in cross-modality fusion backbone are concatenated and reduced in channel, which is called the purpose of *single concatenation*. *Hierarchical concatenation* is to connect these hierarchical features from two modalities in backbone, and adopts the same strategy to reduce dimension. Table II reports the performance of these modules for upscale factor $\times 8$ on CAVE dataset. One can observe that these competitors only collect information from two modalities and do not make full use of the rich RGB content. This makes the results obtained by two networks nearly identical, and their results are relatively poor. In contrast to these modules, the proposed DCFM module modulates RGB features and generates affine transformation coefficients to further enforce deep features in HSI modality, resulting in superior performance. It reveals that our DCFM module can truly improve feature representations compared to typical modules.

E. Ablation Study

The proposed method aims to leverage RGB image to guide the feature learning in the HSI branch. To achieve this end, three key modules are utilized, including DCFM, RDE, and MCPA. To verify whether these modules are effective in improving performance, ablation study is conducted by adding or deleting modules, as shown in Table III. As reported

TABLE IV
PERFORMANCE COMPARISON OF EXISTING APPROACHES ON THREE DATASETS. THE BOLD AND UNDERLINE INDICATE THE BEST AND SECOND PERFORMANCE

Dataset	Upscale factor	Metrics	LTTR	CMS	PZRes-Net	MHF-net	MoG-DCN	UAL	IFMSR
CAVE	×4	PSNR	32.34	34.17	41.78	41.06	43.14	**45.83**	44.64
		SSIM	0.743	0.788	0.953	0.948	0.984	0.990	**0.991**
		SAM	24.28	23.99	12.21	7.85	6.38	5.68	**3.64**
		ERGAS	7.35	6.08	2.57	5.73	2.06	**1.47**	1.80
		RMES	0.024	0.020	0.084	0.009	0.007	**0.005**	0.006
	×8	PSNR	35.41	37.86	41.12	41.35	43.70	**44.76**	43.83
		SSIM	0.890	0.926	0.951	0.959	0.987	0.989	**0.991**
		SAM	19.53	14.27	12.57	6.90	5.74	6.08	**4.25**
		ERGAS	2.46	1.92	1.34	2.59	0.95	**0.81**	0.96
		RMES	0.017	0.013	0.009	0.009	0.007	**0.006**	0.007
Harvard	×4	PSNR	30.61	33.53	39.22	38.05	42.67	40.88	**43.14**
		SSIM	0.689	0.788	0.940	0.922	0.976	0.963	**0.977**
		SAM	17.93	17.86	9.04	8.29	3.83	6.68	**3.35**
		ERGAS	10.21	12.38	5.76	4.23	2.95	3.26	**2.64**
		RMES	0.030	0.022	0.012	0.014	**0.010**	0.011	**0.010**
	×8	PSNR	34.54	37.08	39.06	37.60	**42.84**	42.08	42.64
		SSIM	0.896	0.924	0.946	0.912	**0.977**	0.974	**0.977**
		SAM	12.28	9.92	8.82	10.45	3.66	4.74	**3.53**
		ERGAS	2.90	3.93	2.88	2.14	1.46	1.53	**1.39**
		RMES	0.020	0.015	0.013	0.015	0.010	0.010	**0.009**
Chikusei	×4	PSNR	–	–	36.85	34.08	38.17	36.44	**39.26**
		SSIM	–	–	0.889	0.826	0.934	0.895	**0.951**
		SAM	–	–	12.56	15.67	3.60	**2.96**	12.41
		ERGAS	–	–	8.59	13.90	6.00	7.59	**4.40**
		RMES	–	–	0.015	0.020	0.013	0.016	**0.011**
	×8	PSNR	–	–	35.59	31.77	37.25	35.91	**38.43**
		SSIM	–	–	0.888	0.827	0.935	0.882	**0.954**
		SAM	–	–	12.96	14.12	3.48	**3.28**	12.07
		ERGAS	–	–	4.39	4.20	2.70	4.61	**2.57**
		RMES	–	–	0.017	0.027	0.014	0.017	**0.012**

TABLE V
COMPARISON WHEN TESTING STAGE AMONG MOG-DCN, UAL, AND OUR IFMSR

Method	Parameter	Post-processing	Number of Iterations	Spectral response function
MoG-DCN	12457K	✗	✗	✗
UAL	7098K	✔	1500	✔
IFMSR	1725K	✗	✗	✗

in this table, the values of three evaluation metrics increase markedly when one of the modules is embedded in the network. For instance, to realize the information interaction between two modalities, the RDE module is designed to provide cross-modal contextual guidance and restrain the incompatibility between modalities. Experiments indicate that this module is effective. Notablely, the MCPA module has more parameters than DCFM and RDE modules. This is mainly the parameter produced by aggregating contextual information in the two branches. When any two modules are combined, the experimental results also confirm that these modules are favorable for feature representations, which results in significant performance gains. In particular, the performance gain of the model with RDE and MCPA modules is greatly noticeable. Interestingly, all modules are attached to the model. These values significantly outperform that of any other combinations. The parameters of the model with RDE and MCPA modules are increased by nearly 1000k, compared with single module. These are the main reasons for the significant increase in model parameters. It is concluded that these modules can contribute to the feature learning of the model, thus achieving the desired results.

F. Generalization Ability for Cross Datasets and Scales

To evaluate the performance of the proposed method, this section introduces six approaches to compare generalization ability on different datasets and scales, including LTTR [47], CMS [48], PZRes-Net [24], MHF-net [39], MoG-DCN [40], and UAL [16]. Here, LTTR and CMS are unsupervised, while the other competitors are supervised. Note that LTTR and CMS introduce spectral response function to optimize object function. In particular, UAL incorporates the spectral response function into the loss function to learn the model in the adaptation module. To make a fair comparison, the corresponding loss term is removed on Chikusei with an unknown spectral response function.

Table IV reports the performance comparison for existing approaches on three datasets. Specifically, LTTR and CMS design objective functions in an unsupervised manner, which results in their poor results. In addition, they require the spectral response function to participate in the optimization. This cannot handle HSIs with unknown spectral response functions.

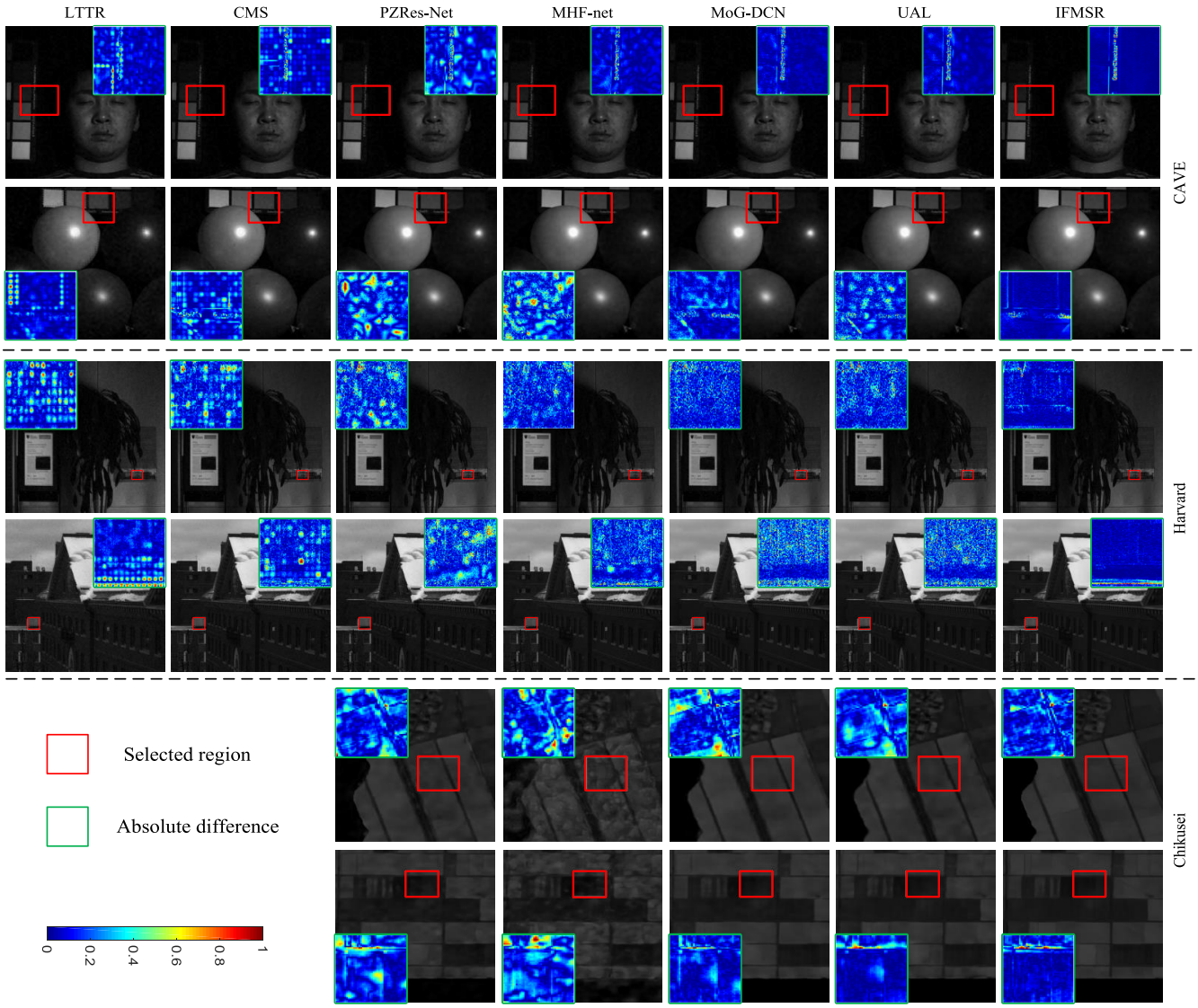

Fig. 4. Visual comparison of spatial reconstruction. The first to three lines represent the visual results of the 15th band, 15th band, and 100th band, respectively.

Therefore, we cannot provide the results on Chikusei dataset. In contrast, the approaches proposed by supervised manner attain the better performance. Among these competitors, MHF-net fails to cope with the image with unknown degeneration. Although PZRes-Net applies parallel input pattern to feed both HSI and RGB images into the model, it lacks the effective use of RGB image with rich details. As a result, its performance is also poor. Overall, UAL, MoG-DCN, and our IFMSR exhibit clear advantages over other competitors in terms of performance. However, UAL and MoG-DCN have more model parameters, as shown in Table V. Moreover, UAL needs to build an independent post-processing using known spectral response function to further optimize the super-resolved results in an unsupervised manner. Importantly, this run takes 1500 iterations, which is surely time consuming. In contrast, the proposed IFMSR has low model parameters and no post-processing. Meanwhile, it can overall address unknown degradation images well on diverse datasets and scales.

We also provide visual results in terms of spatial reconstruction and spectral distortions. To clearly present the contents for spatial reconstruction, the absolute difference of ground-truth and super-resolved result is calculated. Fig. 4 displays the visual comparison by selecting single band in HSI. As can be seen from the figure, the proposed method yields less information in the enlarged region, which accurately indicates that the reconstructed HSI can produce clear edges and textures. The spectral curves in Fig. 5 also reveal that our model is consistent with a ground truth in most cases. Experiments demonstrate the desirable performance of our method in both quantitative and qualitative aspects.

G. Study of Number of CGM-B and CGM-R

CGM-B and CGM-R explore their respective knowledge in cross-modal fusion backbone. To investigate the influence of different numbers of CGM-B and CGM-R, we set the hyper-parameter D to 1–4. Fig. 6 reports the relationship between PSNR performance and parameters. Specifically, as described

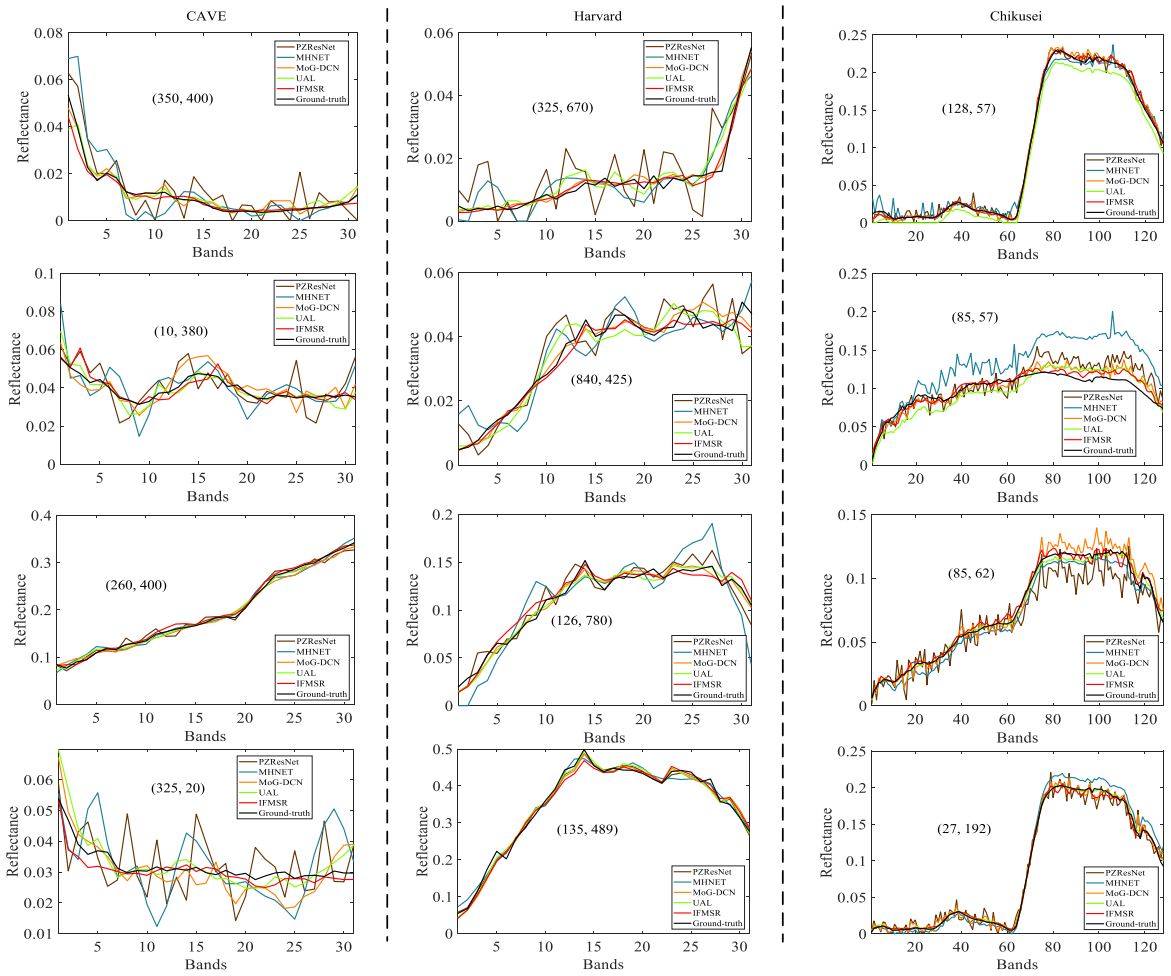

Fig. 5. Visual comparison of spectral distortion for corresponding images by selecting two pixels. (Left to right) Visual results of above images.

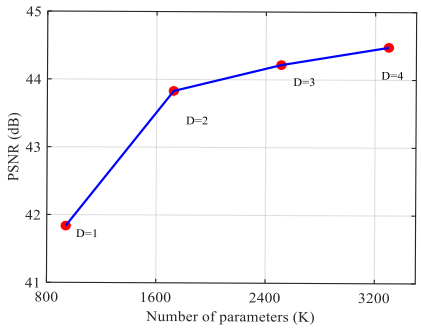

Fig. 6. PSNR performance versus parameter of CGB-B and CGM-R for upscale factor $\times 8$ on CAVE dataset.

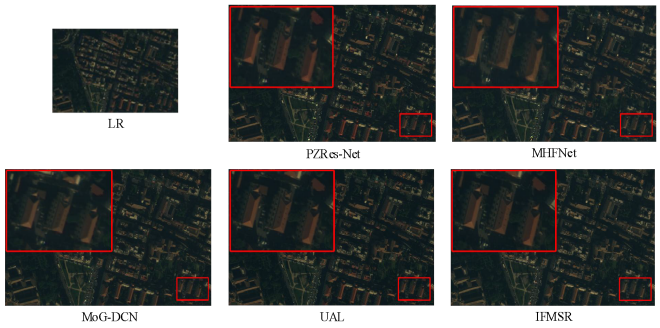

Fig. 7. Visual comparison on real HSI dataset. We choose the 2-3-5 bands after SR to synthesize the pseudo-color image.

in this figure, the hyperparamter D yields a great influence on the model on the whole, especially when its value is set to be minor. As more modules are added, there is a trade-off between the two. Intuitively, the performance gain is not obvious when it comes to $D = 3$. Importantly, this also requires huge computational resources and memory to run. The experiment indicates that both performance and parameters of the model can be well balanced for $D = 2$.

H. Performance on Real HSI Dataset

To examine the applicability of the proposed method, we apply it to a real dataset, Sample of Roman Colosseum, to demonstrate it. Since the HR HSI for this dataset is not available, we utilize the algorithm proposed in [49] to address this issue. Specifically, the HSI is randomly cropped to obtain the patch with size $36 \times 36 \times 8$ and the corresponding RGB image with size $144 \times 144 \times 3$. Then, both images are downsampled by upscale factor $\times 1/4$, respectively. Finally,

TABLE VI
NO-REFERENCE HSI QUALITY MEASUREMENT SCORE ON THE REAL HSI SR TASK

Metric	PZRes-Net	MHF-net	MoG-DCN	UAL	IFMSR
Score	3.68	3.59	3.92	3.36	3.33

the downsampled and original patches are used as training pair. As mentioned above, LTTR [47] and CMS [48] require the spectral response function in their calculations. Hence, the experiments are not shown in this section either. To evaluate the proposed method on real dataset, we introduce nonreference HSI quality evaluation method [50]. Generally, the lower the values mean better visual quality. Table VI displays no-reference HSI quality measurement score on the real HSI SR task. The proposed approach obtains better result. Moreover, Fig. 7 depicts the pseudo-color image on real HSI dataset. We can notice that the proposed method exhibits sharper textures. Meanwhile, more details can be found more precisely in the local enlarged area. This indicates that our IFMSR can handle real-world images, which fully demonstrates its applicability.

V. CONCLUSION

This article proposes an IFMSR to model the dependence of two modalities during feature representation. In this process, each branch assembles high similarity patterns through a global perspective to enhance the exploration of spatial knowledge. Considering that RGB image contains abundant textures, an RDE and a deep cross-modality feature modulation (DCFM) module are designed to supplement and strengthen the feature representations in HSI branch with different depths, generating more edge recovery. Experiment reveals that two modules are effective in performance improvement. Moreover, the results on four datasets demonstrate that our model can achieve comparable performance in terms of quantitative and qualitative evaluation. In the future, we develop the proposed method to add edge information to further guide the model.

REFERENCES

[1] J. Yue, L. Fang, and M. He, "Spectral–spatial latent reconstruction for open-set hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 5227–5241, 2022.

[2] 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, Apr. 2015.

[3] S. Michel, P. Gamet, and M.-J. Lefevre-Fonollosa, "HYPXIM—A hyperspectral satellite defined for science, security and defence users," in *Proc. 3rd Workshop Hyperspectral Image Signal Process., Evol. Remote Sens. (WHISPERS)*, Jun. 2011, pp. 1–4.

[4] X. Zheng, Y. Yuan, and X. Lu, "Hyperspectral image denoising by fusing the selected related bands," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2596–2609, May 2019.

[5] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8693–8703, Oct. 2021.

[6] K. Li, D. Dai, and L. Van Gool, "Hyperspectral image super-resolution with RGB image super-resolution as an auxiliary task," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3193–3202.

[7] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[8] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 7252–7263, 2022.

[9] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "InteractFormer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531715.

[10] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536912.

[11] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

[12] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 63–78.

[13] W. Dong et al., "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.

[14] N. Akhtar, F. Shafait, and A. Mian, "Bayesian sparse representation for hyperspectral image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3631–3640.

[15] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-net for hyperspectral image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2511–2520.

[16] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3073–3082.

[17] 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.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

[18] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 7565–7577, 2020.

[19] M. Yang, L. Jiao, F. Liu, B. Hou, S. Yang, and M. Jian, "DPFL-Nets: Deep pyramid feature learning networks for multiscale change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6402–6416, Nov. 2022.

[20] M. Yang et al., "Coarse-to-fine contrastive self-supervised feature learning for land-cover classification in SAR images with limited labeled data," *IEEE Trans. Image Process.*, vol. 31, pp. 6502–6516, 2022.

[21] X. Han, B. Shi, and Y. Zheng, "SSF-CNN: Spatial and spectral fusion with CNN for hyperspectral image super-resolution," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2506–2510.

[22] 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 Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.

[23] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[24] 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.

[25] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, "Deep blind hyperspectral image fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4150–4159.

[26] J. Hu, T. Huang, L. Deng, H. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[27] K. Jiang et al., "Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining," *IEEE Trans. Image Process.*, vol. 30, pp. 7404–7418, 2021.

[28] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Proces. Syst.*, vol. 27, 2014, pp. 1–9.

[29] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5690–5699.

[30] W. Li, X. Tao, T. Guo, L. Qi, and J. Jia, "MuCAN: Multi-correspondence aggregation network for video super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 335–351.

[31] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3499–3509.
[32] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1604–1613.
[33] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.
[34] E. Wycoff, T.-H. Chan, K. Jia, W.-K. Ma, and Y. Ma, "A non-negative sparse promoting algorithm for high resolution hyperspectral imaging," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 1409–1413.
[35] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5344–5353.
[36] K. Zhang, M. Wang, S. Yang, and L. Jiao, "Spatial–spectral-graph-regularized low-rank tensor decomposition for multispectral and hyperspectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1030–1040, Apr. 2018.
[37] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Nonlocal patch tensor sparse representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3034–3047, Jun. 2019.
[38] X. Han, Y. Zheng, and Y. Chen, "Multi-level and multi-scale spatial and spectral fusion CNN for hyperspectral image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4330–4339.
[39] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1585–1594.
[40] 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.
[41] Q. Wang, Q. Li, and X. Li, "Hyperspectral image superresolution using spectrum and feature context," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 11276–11285, Nov. 2021.
[42] J. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.
[43] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4791–4800.
[44] 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, Sep. 2010.
[45] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. CVPR*, Jun. 2011, pp. 193–200.
[46] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep., SAL-2016-05-27, 2016.
[47] 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, Sep. 2019.
[48] 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, Dec. 2018.
[49] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.
[50] J. Yang, Y. Zhao, C. Yi, and J. C.-W. Chan, "No-reference hyperspectral image quality assessment via quality-sensitive features learning," *Remote Sens.*, vol. 9, no. 4, p. 305, Mar. 2017.

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

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

Maoguo Gong (Senior Member, IEEE) 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, where he was promoted as an Associate Professor and as a Full Professor in 2008 and 2010, respectively, with exceptional admission. His research interests include computational intelligence with applications to optimization, learning, data mining, and image understanding.

Prof. Gong is an Executive Committee Member of the Chinese Association for Artificial Intelligence and a Senior Member of the Chinese Computer Federation. He 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. He is also an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION.

Yuan Yuan (Senior Member, IEEE) 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 coauthored more than 150 articles, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and *Pattern Recognition*, as well as the conference papers in IEEE Conference on Computer Vision and Pattern Recognition, British Machine Vision Conference, IEEE International Conference on Image Processing, IEEE International Conference on Acoustics, Speech and Signal Processing. Her research interests include visual information processing and image/video content analysis.

Qi Wang (Senior Member, IEEE) 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, pattern recognition, and remote sensing.