RAN: Region-Aware Network for Remote Sensing Image Super-Resolution

Baodi Liu, *Member, IEEE*, Lifei Zhao, Shuai Shao, *Member, IEEE*, Weifeng Liu, *Senior Member, IEEE*, Dapeng Tao, *Member, IEEE*, Weijia Cao, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*— The remote sensing (RS) image super-resolution (SR) algorithm aims to reconstruct a high-resolution (HR) image with rich texture details from a given low-resolution (LR) image, improving the spatial resolution. It has been widely concerned in RS image processing and application. Most current deep-learning-based methods rely on paired training datasets. However, most datasets are often based on bicubic degradation. This single construction way limits the performance of the pretrained network. Moreover, SR is an ill-posed problem in that multiple SR images are constructed from a single LR input. This article proposes a region-aware network (RAN) for RS image SR to alleviate the above issues. First, we introduce the contrastive learning strategy to mine the latent degraded representation of the image and serve as the prior knowledge of the network. Considering the RS images are acquired in specific scenes that have apparent self-similarity. Then, we propose a region-aware module (RAM) based on attention mechanisms and the graph neural network to explore region information and cross-patch self-similarity. Extensive experiments have demonstrated that the proposed RAN adapts to RS image SR tasks with various degradations and performs better in constructing texture information.**

***Index Terms*— Attention mechanism, contrastive learning, graph neural network, remote sensing (RS) image super-resolution (SR).**

Manuscript received 20 March 2023; revised 12 June 2023 and 11 October 2023; accepted 21 October 2023. Date of publication 7 November 2023; date of current version 16 November 2023. This work was supported in part by the Yunnan Provincial Major Science and Technology Special Plan Projects under Grant 202202AD080003, in part by the National Natural Science Foundation of China under Grant 61671480 and Grant 62372468, in part by the Major Basic Research Projects in Shandong Province under Grant ZR2023ZD32, in part by the Shandong Natural Science Foundation under Grant ZR2023MF008 and Grant ZR2019MF073, in part by the Qingdao Natural Science Foundation under Grant 23-2-1-161-zyyd-jch, in part by the Major Scientific and Technological Projects of China National Petroleum Corporation (CNPC) under Grant ZD2019-183-008, and in part by the Yunnan Thousand Talents Program under Grant YNQR-CYRC-2019-003. *(Corresponding author: Lifei Zhao.)*

Baodi Liu is with the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China, and also with the State Key Laboratory of Shale Oil and Gas Enrichment Mechanisms and Effective Development, Beijing 100083, China.

Lifei Zhao is with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China (e-mail: upczhaolf@gmail.com).

Shuai Shao is with the Zhejiang Laboratory, Hangzhou 311121, China.

Weifeng Liu is with the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China.

Dapeng Tao is with the School of Information Science and Engineering, Yunnan University, Kunming, Yunnan 650504, China, and also with Yunnan United Vision Technology Company Ltd., Kunming 650299, China.

Weijia Cao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100045, China.

Yicong Zhou is with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China.

Digital Object Identifier 10.1109/TGRS.2023.3330876

I. Introduction

REMOTE sensing (RS) images contain abundant texture details that are crucial for various RS image analysis tasks, including image translation [1], object detection [2], and object tracking [3], [4]. Due to optical conditions and limitations of the sensing equipment, the spatial resolution of RS images often falls short of expectations [5]. RS super-resolution (SR) reconstruction algorithms reconstruct a high-resolution (HR) image with more texture details from the input low-resolution (LR) image, enhancing the spectral quality or spatial resolution of the input image. It provides an effective way for image processing in the field of RS and has attracted widespread attention.

In recent years, deep-learning-based algorithms for RS image SR have achieved significant performance [6], [7], [8]. Most RS image SR algorithms rely on external paired training datasets to improve model performance. This training strategy leads to the pretrained model only performing well in similar degradation. However, the degradation in the real-world application is still being determined. Therefore, it is worth designing the SR module that adapts to various degradation. To alleviate this issue, researchers developed a novel strategy of incorporating the degradation process into the SR network, adjusting the network according to the degradation. SRMDNF [9] is a pioneering work that integrates degraded representations in the reconstructing process. The network extracts the image and degradation information simultaneously. Subsequently, UDVD [10] introduces a dynamic convolution kernel that adaptively enables the network to adjust parameters according to the degradation information. Zhang et al. [11] proposed an alternative strategy that integrates image reconstruction processing and degradation information, enabling the trained module to handle various degraded representations. Later, Hussein et al. [12] introduced correction filters that convert degraded representations within LR images to bicubic degenerates, thereby transforming various types of degradation into bicubic degradation. The ZSSR [13] improves the network's capability to handle diverse degraded representations by incorporating additional estimated blur kernel information during training, which requires thousands of epochs. Subsequently, MZSR [14] employs an optimization-based meta-learning approach to reduce the required iterations. However, these methods are pixel-space degradation estimation that heavily relies on the accuracy of the estimated blur kernel. When there is a discrepancy between the estimated kernel and the true blur kernel, the performance will drastically decrease, resulting in noticeable artifacts in the super-resolved images [15]. Specifically, if the input kernel is smoother than the real one, the output image will be blurry/over-smoothing. Conversely, If the input kernel is sharper than the correct one, the results will be over-sharpened with apparent ringing effects. Contrastive learning strategy has shown potential performance in learning unknown representations [16], [17], [18]. Specifically, it divides existing samples into three categories, namely, query samples, positive samples, and negative samples. It encourages the feature information of query samples to be closer to positive samples while excluding negative samples. Then, obtain the corresponding latent representation of query sample [19], [20], [21]. This strategy estimates degradation in the latent space and alleviates the model's performance decline when the estimated degradation differs from the ground truth [9].

Moreover, RS image SR is an ill-posed problem in that various HR images can be reconstructed according to the input LR image. Many algorithms attempt to reconstruct SR images with the inherent features of LR images to alleviate the above issue, such as local and nonlocal information. Specifically, methods based on local priors [22], such as bilinear or bicubic interpolation, reconstruct unknown pixel values by weighting adjacent pixel values. This approach performs well in nonsaliency areas, i.e., parts of the image with fewer texture details, but results in artifacts in areas with rich texture details, such as edges and texture. Furthermore, influenced by humans' processing of visual information, researchers have proposed attention mechanisms to enable networks to focus limited resources on the most valuable information. Zhang et al. [23] proposed the residual channel attention network (RCAN) based on channel attention mechanisms to address the issue of resource consumption caused by networks treating all channels equally. Meanwhile, Dai et al. [24] introduced second-order statistical information into the attention mechanism and designed the second-order attention network (SAN) to enhance the network's discriminative learning ability for channel information. Although these methods have improved the performance, they only focus on modeling channel attention and ignore nonlocal features.

Methods based on regional nonlocal mean filtering have been proposed to avoid limitations. They search similar image patches over the entire LR image and guide the network for SR reconstruction [25]. Since RS images record specific objects in a particular region, the spectral, texture, and depth feature information in the same scale and different scales have strong similarities and cross-scale patch recurrence [26]. This characteristic is shown in Fig. 1 that multiply green boxes are similar to the yellow. Therefore, it is challenging to explore local and nonlocal prior knowledge to assist the SR reconstruction process. Graph convolutional networks (GCNs) have shown great potential in aggregating data, especially in non-Euclidean spaces [27], [28], [29]. Research has shown that GCN-based networks better extract complex relationships and dependencies in target tasks. This strategy is widely used in image restoration tasks. Zhou et al. [30] construct a dynamic cross-scale feature map by searching for k-nearest neighbor image patches and aggregating them. Liu et al. [26] designed a dual-learning graph convolutional neural network and improved network performance by adding additional constraints.

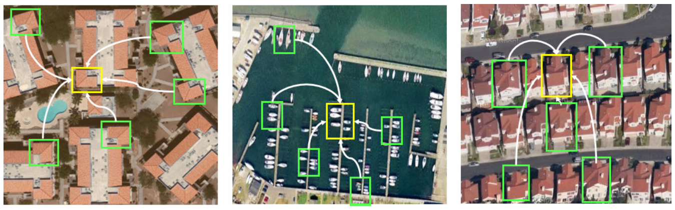

Fig. 1. Illustration of similarities and cross-scale patch recurrence in RS images. Compared with the patches in the yellow box, several similar patches in green exist.

To alleviate the above issues, this article proposes the region-aware network (RAN) for RS image SR. RAN adopts a contrastive learning strategy to obtain the latent degradation representation instead of the specific kernel estimation. Then, designing the region-aware module (RAM) that contains a 3-D attention module that models channel and spatial information, and a cross-patch aggregation module based on GCN to explore nonlocal information. RAM motivates the network to focus on local and nonlocal information. The main contributions can be summarized as follows.

1) This article proposes a RAN that dynamically adapts to RS SR tasks with different degradation and reconstructs SR images with enriched texture details. Extensive experiments on RSC11, UC-Merced, and NWPU45 datasets have demonstrated the validity and superiority of performance.
2) This article presents a degradation-aware module based on the contrastive learning strategy to learn latent degradation representation within the image. This implicit representation-based degradation module alleviates the problem of the model performance degradation caused by degradation mismatch.
3) This article designs a RAM, which includes a 3-D attention mechanism and a cross-patch feature aggregation module (CPFAM). The 3-D attention mechanism models channel-wise and spatial features to obtain the weighted feature map. The cross-patch aggregation module based on the graph convolutional neural network incorporates the cross-scale self-similarity within the image.

II. Related Work

A. RS Image SR

RS image SR algorithms based on convolutional neural networks have demonstrated powerful feature extraction capabilities and outstanding performance, which have been widely applied [31].

Liebel and Körner [32] introduced convolutional neural networks to RS image SR tasks as a pioneering work. The fine-tuned convolutional neural network is applied to multispectral remote-sensing images. However, this method only focuses on the performance of the third band of multispectral images. To address this issue, Tuna et al. [33] applied the

VDSR [34] that models the intensity-hue-saturation (IHS) transformation of RS images and obtain twofold, threefold, and fourfold VHR SPOT6&7 and Pleiades 1A&B RS images. Based on this method, an RS deep residual learning network (RS-DRL) is proposed. Relevant experiments are designed to demonstrate that RS-DRL outperforms the VDSR method on the Sentinel-2A dataset. However, these methods ignore complex internal features of RS images and directly apply image SR algorithms in the computer vision field to the RS field. This strategy limits the expression performance of remote-sensing image SR models. Therefore, designing RS image SR algorithms that adapt to the RS field has received widespread attention.

Lei et al. [35] designed a local-global combined network (LGCNet) that obtains multilevel feature representations of RS images by connecting features from different convolutional layers. This architecture encourages the network to learn detailed local and global features, including edges, contours, and environmental information. With ample information, the network reconstructs high-quality remote-sensing images. In addition, researchers proposed upsampling and downsampling units to save computational resources consumed by the algorithm. To further improve the performance, many researchers have proposed network structures based on dense residual modules. Jiang et al. [36] proposed a deep distillation recursive network (DDRN), which includes ultradense residual blocks (UDBs) and multiscale purification units (MSPUs) as well as a distillation mechanism. MSPU compensates the high-frequency components but loses in the feature information transmission process. DDRN performs well on the Jilin-1 video RS dataset. Subsequently, Deeba et al. [37] proposed a wide RS residual (WRSR) network, which gradually reduces the depth as the width of the residual network increases. This training strategy enhances the sensitivity of the network to the loss function and improves network performance.

The above RS image SR algorithms only apply the network to upsample LR images to obtain HR images, ignoring the value of the LR image's internal features. Zhang et al. [38] proposed a progressive residual depth neural network (PRDNN) that learns different level features through various receptive fields. Based on multiscale information, PRDNN reconstructs finer edge and texture information. Based on PRDNN, Shao et al. [39] proposed a coupled sparse auto-encoder (CSAE) that utilizes sparse learning. Then, the obtained prior knowledge is treated as prior knowledge, enabling the network to learn more accurately the relationship between LR and HR images.

The SR task is a cross-scale task that reconstructs an HR image from the input LR image. Therefore, it is worth exploring a more effective way to learn internal features. And reasonably allocate limited resources to different features and further alleviate the ill-posed problem.

B. Attention Mechanism

The human biological system focuses limited sources on the most valuable information when processing massive resources. This process allocates attention and resources reasonably and improves effectiveness and accuracy. Inspired by this attention mechanism, many researchers have explored and successfully applied attention mechanisms to RS image SR reconstruction tasks. Gu et al. [40] proposed a deep RCAN, which models the interdependence between channels through local features and channel attention mechanisms, enhancing the network's feature representation ability. Dong et al. [41] proposed a multiperception attention network (MPSR) composed of an improved residual network and a channel attention mechanism. Peng et al. [42] added gated convolutional units with long skip connections based on channel attention mechanisms to enhance the network's attention to images' high-frequency information.

However, the above methods only introduce channel attention mechanisms into the network or each residual block, which makes it difficult for the network to learn deeper features of LR images. To address this issue, Dong et al. [43] proposed a second-order multiscale SR network. Li et al. [44] designed a fused recurrent network, which utilizes a channel attention mechanism to preserve and fuse high-order local features of both low and HR images. The aforementioned methods only consider the information from RS images with different spatial resolutions, ignoring the relationship between different scenes' information.

To address this, Zhang et al. [45] proposed a multiscale attention mechanism to extract multilevel information in different scenes and enhance the network's feature learning ability. Huang and Jing [46] designed a dual-attention module, which includes a local multilevel fusion mechanism and a dual-attention mechanism. This model structure makes the network pay more attention to high-frequency information. Based on the dual-attention module, the authors design a deep residual dual-attention network to fuse global and local information. However, this method only combines different hierarchical information of the same dimension, such as different channel features, ignoring the fact that different hierarchical information of different features is effective for the network. Wang et al. [47] proposed a nonlocal up–down convolution attention mechanism network, including a nonlocal feature enhancement module, enhanced upsampling channel attention, and an enhanced downsampling spatial attention mechanism module. This algorithm improves the quality of the reconstructed image by fusing nonlocal feature information, channel information, and spatial information. RS images are recorded RS elements with specific area, spectral information, texture features, and depth features within the image. They have strong characteristics of smoothness and self-similarity at the same scale or different scales. Therefore, it is worth designing an effective model to enhance the network's attention on local and nonlocal features.

III. Methodology

This section first demonstrates the learning strategy of degradation representation, then illustrates the overall network framework of the proposed RAN for RS image SR reconstruction, and then introduces the detailed structure of the proposed RAM.

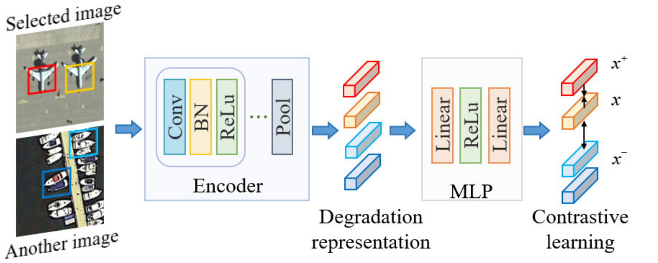

Fig. 2. Degradation representation module based on contrastive learning. The yellow patch is the query sample, the red inside the same image is a positive sample, and the blue patches in other images are negative. Assuming that the degradation representations in the same image are the same, the degradation representations in different images are different. An encoder encodes the above three types of samples, and x, x^+, and x^- are obtained through the MLP layer.

A. Degradation Representation

The degradation model of LR RS images can be expressed as follows:

$$I_{\mathrm{LR}} = (I_{\mathrm{HR}} \otimes k) \downarrow_s + n. \tag{1}$$

Among them, I_{HR} is an HR image, k is a blur kernel, $\otimes$ denotes a convolution operation, $\downarrow_s$ represents a s times downsampling operation, and n is the additional Gaussian noise added.

According to the way that the model handles the degradation, the existing RS SR algorithms can be divided into two categories: first, the model is trained on constructed datasets with specific degradation, such as bicubic degradation. Such a pretrained model merely performs well in similar degraded situations. If the degradation in test images is unrelated, the performance will apparently decline. Second, the methods use the network to estimate the degraded representation and get the specific blur kernel information. This information is input into the network as prior knowledge. In this way, the performance largely depends on the accuracy of the estimated blurry kernel. The model performance will drop sharply when the estimated result mismatches the real kernel. In addition, there is a domain gap between image and degradation information. Thus, directly encoding both in the same convolutional network will interfere with results.

The contrastive learning strategy learns the latent features of the target sample in an unsupervised manner [48]. It has shown superior maximizing mutual information in the latent space and learning unknown representations. Specifically, in contrastive learning, the existing samples are divided into three categories: query, positive, and negative. The feature information of query samples is encouraged to be closer to positive samples while excluding negative samples. With the loss function, the potential representation of query samples is obtained. Inspired by this, this article introduces the contrastive learning strategy to mine the degradation representation.

The degradation representation based on contrastive learning is shown in Fig. 2. The patch of yellow is a query sample, another red in the same image is a positive sample, and two image patches of blue in another image are negative samples. Suppose the query sample is x, the positive sample is x^+, and the negative samples are x^-. Three types of patches are encoded as corresponding degradation representation and further input into the multilayer perceptron (MLP), then measure the similarity through the InfoNCE loss function as follows [49]:

$$L_x = -\log \frac{\exp\left(x \cdot x^+/\tau\right)}{\sum_{n=1}^{N} \exp\left(x \cdot x_n^-/\tau\right)} \tag{2}$$

where N denotes the total number of negative samples, τ indicates a temperature hyper-parameter, and $\cdot$ represents the dot product between two vectors.

B. Region-Aware Network

Currently, deep-learning-based RS image SR methods rely on synthetic training datasets to improve performance. The constructed way is based on bicubic downsampling or another simple mode. Compared with the realistic situation, this degradation is too idealistic. It results in pretrained models only performing well on similar degradation. To alleviate this issue, this article designs the RAN based on degradation representation. The RAN is shown in Fig. 3. Specifically, RAM denotes a RAM, which is demonstrated in Section III-C. And DR represents the obtained degradation representation. Then, it passes through two fully connected (FC) layers as the input of the RA-Conv module. One branch passes through a reshape operation and obtains the convolutional kernel $w \in R^{C\times 1\times 3\times 3}$. Then, the input feature F through the convolutional layer with w, ReLU layer, and 1×1 convolutional layer to obtain this branch's output feature F_1. Inspired by interactive image restoration tasks, convolutional neural networks adaptively adjust feature maps to handle different types of degradation by changing input degradation representation [50]. Therefore, RAN takes the degradation representation and generates channel modulation coefficients v after passing through a Sigmoid layer, which scales the channel information as F to obtain the scaled feature F_2. Finally, adding features F_1 and F_2, the above series operations are represented as RA-Conv. Then, with the input feature, the output feature F_{out} is obtained through the ReLU layer, convolutional layer, ReLU layer, RA-Conv, ReLU layer, and convolutional layer. The above operation is denoted as RA block (RA-B). The residual group (RG) contains 5 RA-Bs, and after 5 RGs, with an upsampler module, RAN obtains the SR image I_{SR}. This structure enhances the adaptability to different degradation processes and improves information interaction between the front and end of the network.

C. Region-Aware Module

The attention mechanism is widely applied in RS SR procedures as an effective module to improve network performance. However, remote-sensing images contain complex texture information, structural features, and various scenes. The learning strategy that only enhances a single channel or spatial feature limits the performance. Therefore, we designed a RAM in this section, mainly containing two parts. First, introduce a 3-D attention module, which obtains the feature $\tilde{F}$ by combining channel and spatial attention. Second, we design a CPFAM based on a graph neural network and obtain the feature F_q. Then, connecting the weighted feature map $\tilde{F}$ in

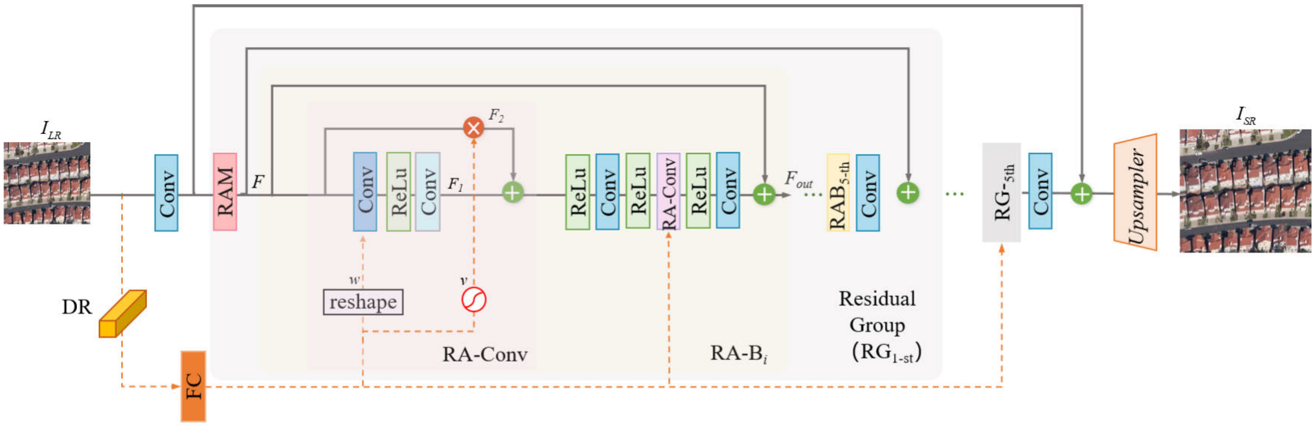

Fig. 3. Overall structure of RAN. RAM is the proposed RAM. RA-Conv adapts the features based on the degradation representation, which predicts the kernel w and adjusts channel-wise according to the coefficient v.

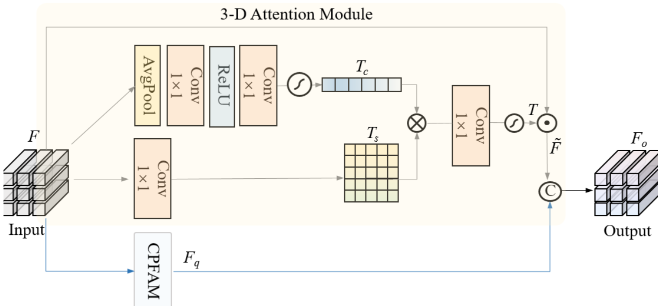

Fig. 4. Illustration of RAM. RAM includes two branches: a 3-D attention module and a CPFAM. The former branch extracts the weighted feature map and local information. For the latter, the branch extracts patch-wise consistency and nonlocal information. ⊗ is matrix multiplication, ⊙ is element multiplication, C denotes the concatenate operation.

the 3-D module and aggregation feature F_q, as the output F_o. The illustration is shown in Fig. 4.

1) Three-Dimensional Attention Module: Assuming the feature map input to the 3-D attention module is $F \in \mathbb{R}^{C\times H\times W}$, the weighted channel feature is obtained through the average pooling layer, convolutional layer, ReLU activation function, convolutional layer, and sigmoid layer, as shown in (3). The weighted spatial attention feature is obtained by convolving the input feature map, as shown in (4). Then, the above matrices are multiplied by a convolutional layer and a sigmoid layer to obtain the 3-D attention-weighted feature. Finally, element-wise addition of the weighted feature map and the input feature map F yields the feature map $\tilde{F}$, as shown in (5)

$$T_c = \mathrm{Sigmoid}\left(f^{1\times1}\left(\mathrm{ReL}\,U\left(f^{1\times1}(\mathrm{Avg}(F))\right)\right)\right) \quad (3)$$

$$T_s = \mathrm{Sigmoid}\left(f^{1\times1}(F)\right) \quad (4)$$

$$\tilde{F} = \left(\mathrm{Sigmoid}\left(f^{1\times1}(T_c \otimes T_s)\right)\right) \odot F \quad (5)$$

where Avg is global average pooling, $f^{1\times1}$ is a convolution operation with a convolution kernel of 1 × 1, ⊗ is matrix multiplication, and ⊙ is element multiplication. The 3-D attention mechanism retains the spatial features while learning the channel feature information, which is beneficial for capturing the internal area information of the image. This structure enhances the 3-D information expression and strengthens the learning and discrimination ability of the network.

2) Cross-Patch Feature Aggregation Module: Many nonlocal methods demonstrate outstanding aggregation performance in SR tasks due to their effective ability to model long-range dependencies [50]. The nonlocal aggregation method in deep neural networks can be represented as

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right) \quad (6)$$

where x_i is the ith of the input x (e.g., an image, sequence, video, and other features), x_j represents the jth neighbor of x_i. $g(x_j)$ is the feature value of x_i, $f(x_i, x_j)$ denotes the weight of the aggregated feature $g(x_j)$, and $C(x)$ is the normalization factor.

Although convolution operation demonstrates superior performance in capturing local features, it is limited to modeling spatially irregular image patches' features. With the kernel size increase, the dependence between long-distance features also cannot be effectively maintained. Graph convolutional neural networks maintain the invariance relationship between nodes during the modeling process. RS images are obtained under specific conditions, with rich cross-scale similarity within the image. Therefore, this article designs a CPFAM, as shown in Fig. 5. It regards patches and similarity as nodes and edges. And explores self-similarity features based on image patches while maintaining spatial relationships.

The primary process of the CPFAM contains graph construction and feature aggregation processes. Based on cross-scale patches' self-similarity, the graph convolutional neural network captures long-range dependency relationships between cross-scale image patches. The graph construction procedure is conducted in Algorithm 1. SR is a cross-scale task in that LR images are reconstructed into HR images through the pretrained network. Thus, cross-scale information within images effectively guides the reconstruction process.

First, downscaling the input image I_{LR} to obtain downsampled image $I_{\mathrm{LR}\downarrow_s}$. The embedding features of the downsampled image are extracted from the first three layers of the VGG-19 network [51], denoted as $E_{\mathrm{LR}\downarrow_s}$. The size of the sliding window is $d \times d$, searching k-nearest neighboring patches of the patch

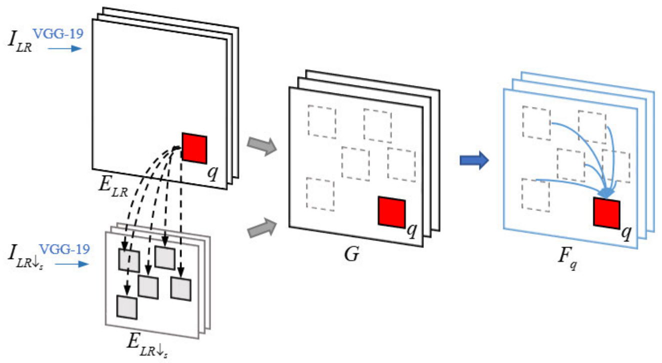

Fig. 5. Illustration of the CPFAM, which contains graph construction and feature aggregation processes. Among them, I_{LR} is the input image, $I_{\mathrm{LR}\downarrow_s}$ is the downsampled image, the E_{LR} and $E_{\mathrm{LR}\downarrow_s}$ are corresponding embedding feature, q is a query feature patch in E_{LR}, G is the constructed graph, and F_q is the aggregated feature.

Algorithm 1 Graph Construction

```
Input: Input image $I_{LR}$, Feature $E_{LR}$
Output: Graph neural network $G$, the Euclidean
        distance $D_{n_{i\to q}}(i = \{1, \ldots, k\})$
Downsampling the input image $I_{LR}$ as $I_{LR\downarrow_s}$.
Obtaining the embedding feature of $I_{LR\downarrow_x}$ through
  VGG-19 network, denoted as $E_{LR\downarrow_s}$.
Finding $k$-nearest neighbor of $q$ with size $d \times d$ in
  $E_{LR\downarrow_s}$, denoted as $E^i_{LR\downarrow s}$, along with their
  corresponding indicates $index_i(i = \{1, \ldots, k\})$.
Calculating the Euclidean distance between $E^q_{LR}$ and
  $E^i_{LR\downarrow s}$, which is formulated as: $D_{n_{i\to q}} = E^q_{LR} - E^i_{LR\downarrow s}$.
According to the $index_i$, corresponding $k$ patches with
  $ds \times ds$ in $E_{LR}$ as $E^i_{LR}$.
return $index_i(i = \{1, \ldots, k\})$, $D_{n_{i\to q}}$
```

q in $E_{\mathrm{LR}\downarrow_s}$. The difference vector is calculated based on the Euclidean distance, as $D_{n_{i\to q}} = E^q_{\mathrm{LR}} - E^i_{\mathrm{LR}\downarrow_s}$. Then, mapping patches to E_{LR} as E^i_{LR}. Finally, a graph convolutional neural network G is constructed, which $E^i_{\mathrm{LR}\downarrow s}(i = \{1, \ldots, k\})$ and E^q_{LR} as vertices, $D_{n_{i\to q}}$ as edges.

The edge-conditioned convolution [52] is proposed to adaptively acquire the weight values of each target's neighbor patches. This article aggregates the features of k image patches in the graph G, which can be formulated by

$$F_q = \frac{1}{\delta_q(F_{\mathrm{LR}})} \sum_{i \in \mathcal{S}_q} \exp\big(\mathrm{ECC}\big(\mathcal{D}_{n_i \to q}\big)\big) F^i_{\mathrm{LR}} \tag{7}$$

where exp($\dot{}$) is the exponential function that accelerates and stabilizes the training process. S_q represents the collection of k-nearest neighboring patches of the target q. $\delta_q(F_{\mathrm{LR}})$ is the normalization factor that maintains the stability and robustness of the process. The cross-scale image patch aggregation features F_q is obtained through Algorithm 2.

D. Discussion

The RAN aims to alleviate performance decline when the pixel-level estimate result mismatches the actual. RAN adopts a contrastive learning strategy to obtain the latent degradation representation within the input. Moreover, RS images with strong characteristics of self-similarity across patches. Based on the attention mechanism and graph neural network, this article designs a RAM that motivates the network to model the channel and spatial and cross-scale features.

Algorithm 2 Feature Aggragation

```
Input: Feature map $F$, indication of $k$ nearest
       neighboring index $index_i(i = \{1, \ldots, k\})$, the
       Euclidean distance $D_{n_{i\to q}}(i = \{1, \ldots, k\})$
Output: The aggregation feature $F_q$
Obtaining corresponding feature $k$ nearest neighbor
  patches' feature $F^i_{LR}$ according to the indication
  $index_i(i = \{1, \ldots, k\})$.
Calculating aggregation feature $F_q$ according to Eq. 7.
return $F_q$
```

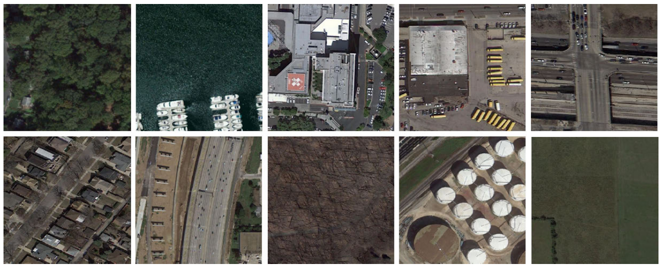

Fig. 6. Selected image of RSC11 dataset.

IV. Experiments

This section introduces the benchmark datasets briefly and demonstrates the experimental implementation in detail. Then, we demonstrate and discuss the experimental results. Next, we design ablation experiments to analyze the performance-influenced factors. All experiments are conducted on a Tesla-V100 GPU with 32G memory.

A. Datasets

The 800 training images in the DIV2K dataset are selected as the training set. All comparisons are conducted on the RS RSC11 dataset, UC-Merced dataset, and NWPU-RESISC45 dataset to verify the performance.

The DIV2K dataset [53] is a high-quality dataset, which is used in the NTIRE 2017 example-based single-image SR competition. It contains 1000 images, including 800 images in the training set, 100 in the validation set, and 100 in the test set. It contains a wide range of content, including people, handcrafted items, urban and rural environments, plants, animals, underwater natural landscapes, and natural landscapes under low-light conditions.

The RSC11 dataset [54] is a collection of RS images manually extracted from Google Maps, covering 7 areas and containing 11 complex scene categories, e.g., dense forests, grasslands, and ports. The dataset consists of 1232 images. Each category contains 100 images with 512 × 512 pixels. Some selected images are shown in Fig. 6.

Fig. 7. Selected image of UC-Merced dataset.

Fig. 8. Selected image of NWPU45 dataset.

The UC-Merced dataset [55] is manually extracted from the National Map Urban Area Imagery collection of the U.S. Geological Survey, covering 20 cities and containing 21 scene categories, e.g., agriculture, baseball diamond, and dense residential. Each category contains 100 images of 256 × 256 pixels. Some selected images are shown in Fig. 7.

The NWPU45 dataset [56] contains a total of 31 500 images covering 45 scene categories, with 700 images per category, e.g., commercial area, dense residential, and wetland. The size of each image is 256 × 256 pixels. Some selected images are shown in Fig. 8.

The LR images are synthesized according to (1). The kernel widths σ are set to [0.2, 2.0] and [0.2, 4.0] for ×2 and ×4 SR tasks, respectively.

B. Implementation Details

During training, this article adopts the Adam optimizer (with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.09$). The data augmentation is performed through random rotation and flipping. Then, this article randomly selects 32 Gaussian kernels from the above ranges to generate corresponding LR images. The training process consists of 600 epochs, during which the encoder is degraded in the first 100 iterations. The initial learning rate is set to 0.001 and decreased to 0.0001 after 60 iterations. The loss function is defined as (2), $N = 8192$, $\tau = 0.07$, size of x^+, x^-, x is set as 48 × 48. The encoder and SR network are optimized in the following 500 iterations. The initial learning rate is 0.0001 and decreases by half every 125 epoch. The loss function is defined as L_1 loss between the super-resolved image and the ground truth HR image, along with

$$\begin{aligned} L &= L_1 + L_x \\ &= \sum_{i=1}^{N} L_1\left(\mathrm{RAN}\left(I_{\mathrm{LR}}^i\right), I_{\mathrm{HR}}^i\right) + L_x. \end{aligned} \tag{8}$$

Then, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [59] are selected as evaluation indexes. PSNR measures the quality of reconstructed images, as shown in (9). SSIM compares luminance, contrast, and structure between I_{SR} and I_{LR}, and the formulation is represented in (10)

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(I_{\mathrm{HR}}(i) - I_{\mathrm{SR}}(i))^2}\right) \tag{9}$$

where L represents the maximum value of the image pixel as 255. The larger the value of PSNR, the better the visual effects

$$\mathrm{SSIM} = [\mathcal{C}_l(I_{\mathrm{HR}}, I_{\mathrm{SR}})]^{\alpha}[\mathcal{C}_c(I_{\mathrm{HR}}, I_{\mathrm{SR}})]^{\beta}[\mathcal{C}_s(I_{\mathrm{HR}}, I_{\mathrm{SR}})]^{\gamma} \tag{10}$$

where $\mathcal{C}_l$, $\mathcal{C}_c$, and $\mathcal{C}_s$ denote the luminance comparison, contrast comparison, and structure comparison, respectively. The range value of SSIM is from 0 to 1. The higher, the better quality.

C. Experimental Results

1) Quantitative Results: This article compares the proposed RAN method with recent SR algorithms, including RCAN [23], SRMDNF [57], MZSR [14], IKC [58], and DASR [9]. The experimental results are shown in Table I. The best result is indicated by "**bold**" and the second-best result is shown in "underline."

Among them, RCAN currently has the highest PSNR evaluation metric on bicubic degradation. MZSR is based on meta-learning and can handle any Gaussian blur kernel. SRMDNF introduces blur features, enabling the network to adapt to different Gaussian kernels and noise. IKC continuously adjusts parameters based on the SR results, adapting the network to different degraded SR tasks. DASR is a prior work that introduces a contrastive strategy to SR tasks. All comparisons are conducted on publicly available pretrained models.

For the ×2 SR task, we adopt kernel widths 0, 0.6, 1.2, and 1.8, while for the ×4 SR task, we adjust kernel widths as 0, 1.2, 2.4, and 3.6. As the kernel width increased, the images became more blurry. From Table I, it can be seen that the RCAN method performed best when the kernel width was 0, which represents bicubic degradation. The SR performance sharply decreased as the kernel width increased, and the degradation deviated further from bicubic degradation. Although the SRMDNF and MZSR methods can handle images with different degradation representations by estimating the blur kernel, the results show that these two methods depend on the accuracy of the estimated blur kernel. It limits the expression of the model's performance when there is a deviation in the estimated kernel. The IKC method's continuously corrected blur kernel strategy performs better than the above three methods, but the correction process consumes a lot of computational resources. Although DASR adapts to various degradation, it ignores

TABLE I

QUANTITATIVE COMPARISON RESULTS ON RSC11, UC-MERCED, AND NWPU45 DATASETS. BOLD AND UNDERLINE INDICATE THE BEST AND THE SECOND-BEST PERFORMANCE, RESPECTIVELY

Method	**Scale**	**RSC11**				**UC-Merced**				**NWPU45**			
		0	**0.6**	**1.2**	**1.8**	**0**	**0.6**	**1.2**	**1.8**	**0**	**0.6**	**1.2**	**1.8**
Bicubic	×2	30.55	30.54	28.36	26.73	31.74	31.64	28.82	26.65	31.38	31.31	28.92	27.07
RCAN [23]		**33.68**	30.69	29.07	26.85	**36.03**	33.45	30.17	26.99	**34.16**	33.12	30.09	27.33
SRMDNF [57]		23.29	23.26	23.15	22.99	22.38	22.33	22.18	21.97	23.15	23.11	22.97	22.78
MZSR [14]		29.46	29.24	28.47	26.99	30.21	30.18	29.56	28.97	29.75	29.52	28.96	27.41
DASR [9]		32.06	31.78	30.75	30.04	34.72	33.89	31.33	30.69	32.86	32.17	31.46	31.04
RAN (Ours)		32.23	**32.13**	**32.12**	**31.02**	34.93	**34.17**	**33.84**	**32.38**	33.73	**33.18**	**33.07**	**31.79**

Method	**Scale**	**RSC11**				**UC-Merced**				**NWPU45**			
		0	**1.2**	**2.4**	**3.6**	**0**	**1.2**	**2.4**	**3.6**	**0**	**1.2**	**2.4**	**3.6**
Bicubic	×4	26.48	26.45	25.41	24.45	26.69	26.68	24.81	23.48	26.90	26.86	25.52	24.35
RCAN [23]		**27.98**	26.76	26.11	24.64	**29.14**	27.15	25.81	23.76	**28.68**	27.33	26.27	24.52
SRMDNF [57]		21.46	21.44	21.37	21.27	19.97	19.95	19.87	19.75	20.89	20.87	20.79	20.68
IKC [58]		26.56	26.63	26.54	25.87	27.56	27.60	27.42	26.41	27.25	27.29	27.15	26.42
DASR [9]		27.16	26.87	26.06	25.88	27.73	27.71	27.43	26.43	27.55	27.27	27.16	26.51
RAN (Ours)		27.96	**27.71**	**27.59**	**26.95**	28.43	**28.38**	**28.27**	**27.48**	28.29	**28.33**	**28.23**	**27.48**

the internal information of RS images. The proposed RAN adopts a contrastive learning strategy to estimate the implicit degradation representation and enable the network to adapt different kernel width degradations within images. In addition, the RAM explores the channel and spatial information with 3-D attention and captures cross-scale similarity. It assists the network in the reconstruction process. Thus, RAN has the best robustness performance on different datasets and degradations.

Compared with suboptimal results, the proposed RAN achieves PSNR improvements of 0.35, 1.37, and 0.98 dB with kernel widths of 0.6, 1.2, and 2.4, respectively, with a scale factor of 2 on the RSC11 dataset. For a scaling factor of 4 and kernel widths of 1.2, 2.4, and 3.6, the PSNR was improved by 0.84, 1.53, and 1.07 dB, respectively. On the UC-Merced dataset, the PSNR was improved by 0.28, 2.51, and 1.69 dB with kernel widths of 0.6, 1.2, and 2.4 with a scale factor of 2, respectively. For a scale factor of 4 and kernel widths of 1.2, 2.4, and 3.6, the PSNR was improved by 0.67, 0.84, and 1.05 dB, respectively. On the NWPU45 dataset, the PSNR was improved by 0.06, 1.61, and 0.75 dB with kernel widths of 0.6, 1.2, and 2.4 with a scale factor of 2, respectively. For a scale factor of 4 and kernel widths of 1.2, 2.4, and 3.6, the PSNR was improved by 1.04, 1.07, and 0.97 dB, respectively. Compared with suboptimal methods, the proposed RAN method achieves a PSNR improvement of approximately 1 dB on the above RS datasets. Although the SR performance of RAN also declines with increasing kernel width, satisfactory SR results could still be obtained. These experimental results demonstrate the effectiveness and superiority of the RAN method.

2) Qualitative Results: The visualization results are shown in Figs. 9 and 10, which present the qualitative comparison of the RSC11, UC-Merced, and NWPU45 datasets. The compared methods adopt publicly available pretrained models. The σ denotes the width of the isotropic Gaussian kernel, the wider, and the more blurry. It is noted that $\sigma = 0$ represents the bicubic degradation. As the kernel width increases, the SR performance of Bicubic, RCAN, SRMDNF, IKC, and RCAN all decline and reconstruct blurred SR images. The results of RAN have more evident shapes and edges, resulting in the best visual effects of the SR images.

Bicubic is based on interpolation and utilizes linear operations to obtain the pixel values at the target position without any additional information. Deep-learning-based methods, such as RCAN, SRMDNF, MZSR, and RAN, utilize convolutional neural networks to learn embedded features inside the image and infer some texture details of the SR image. However, due to the limitations of feature learning and the failure to fully utilize the learned features, these methods produce blurred contours and artifacts. Although the RCAN method generates relatively clear texture details at $\sigma = 0$, the performance sharply declines when the degradation representations differ from bicubic. It results from training datasets are based on bicubic degradation. Consequently, RCAN often generates images with obvious artifacts.

SRMDNF and MZSR rely on estimated degradation representation, which limits the model's performance. When the estimated degradation representation is similar to the actual degradation representation, they generate clearer SR images. However, they produce blurry SR images when the estimated degradation deviates from the actual degradation. IKC improves the visual effect by continually adjusting the estimated degradation representation during reconstruction. However, there exists an inherent interdomain between the degradation and images, which leads to blurry or ringing effects in SR images. DASR adaptively constructs SR images with degradation representation. However, it ignores the internal information and characteristics of remote-sensing images. The RAN method models the channel and spatial features and combines graph neural networks to explore the self-similarity of images. The performance of this method is superior and

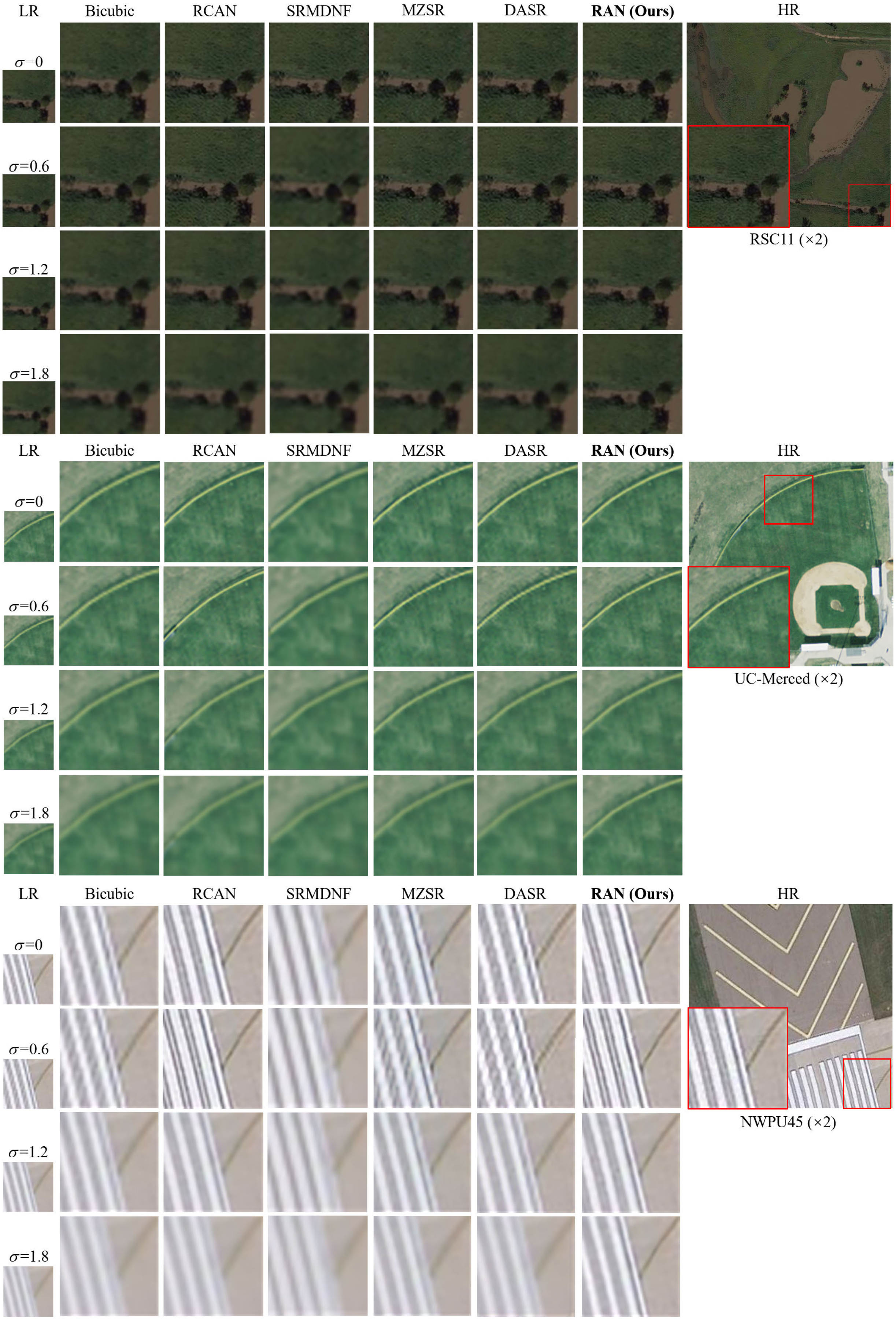

Fig. 9. Qualitative comparison of the scale factor ×2 on the RSC11, UC-Merced, and NWPU45 datasets.

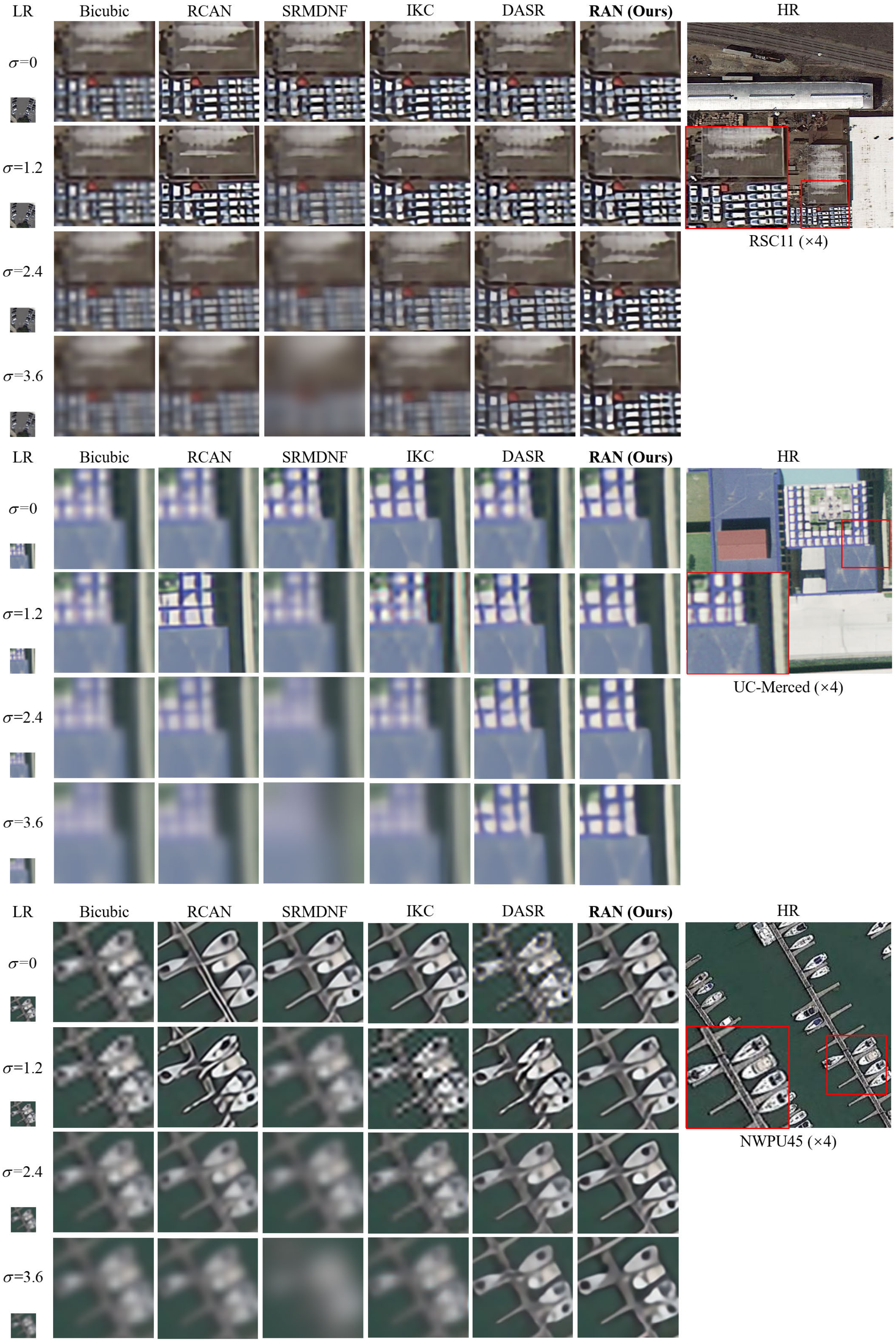

Fig. 10. Qualitative comparison of the scale factor ×4 on the RSC11, UC-Merced, and NWPU45 datasets.

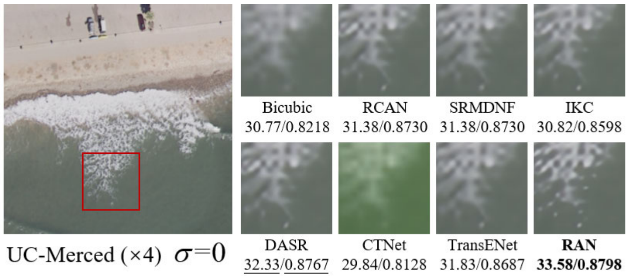

Fig. 11. Qualitative comparison with recent RS SR methods on ×4 UC-Merced dataset ($\sigma = 0$).

TABLE II

Ablation study of Degradation Representation and RAM on ×4 $\sigma = 3.6$ NWPU45 Dataset

Degradation Representation	Region-aware Module 3-D Attention	 Cross-Patch Feature Aggregation	PSNR
	✓	✓	24.58 (⇓ 2.9dB)
✓		✓	26.87 (⇓ 0.61dB)
✓	✓		26.53 (⇓ 0.95dB)

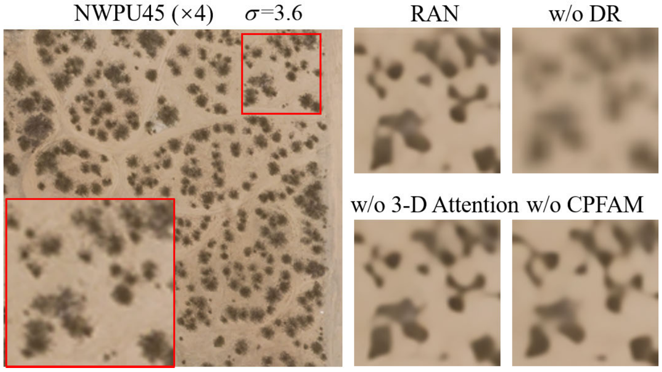

Fig. 12. Visualization results achieved on ablation study of ×4 $\sigma = 3.6$ NWPU45 dataset.

generates SR images with more apparent texture details and better visual quality.

Moreover, we further compare RAN with recent RS image SR algorithms, including CTNet [7] and TransENet [8]. CTNet focuses on hierarchical feature learning, which extracts feature representation and enhances performance through context feature transformation. TransENet is based on the transformer structure, which leverages and fuses multilevel features. Both methods are designed for bicubic degradation. The comparison is conducted on ×4 UC-Merced dataset ($\sigma = 0$) case, and the qualitative result is shown in Fig. 11. Compared with other methods, the proposed RAN reconstructs more texture details and achieves the best quantitative results.

D. Ablation Studies

In this section, we design ablation studies on the NWPU45 dataset to analyze the efficiency of our methods block by block, including degradation representation and RAM. The detailed results are listed in Table II and visualization results are shown in Fig. 12.

The RAN benefits from the degradation representation to present discriminative degradation knowledge. This article designs a network variant to prove its contribution by canceling the degradation representation. Compared with RAN, the evaluation index declines by 2.9 dB. The result has demonstrated that degradation representation plays a crucial role in handling SR tasks with degradation information. Without degradation representation, SR images often contain obvious blurry artifacts. Then, this article removes the 3-D attention in the CPFAM. Compared with the RAN, the PSNR has declined 0.61 dB. The 3-D attention module effectively models channel and spatial attention, which enhances feature interaction. Without the 3-D attention module, some details are not well restored. Last, the CPFAM is removed, and the PSNR has decreased by 0.95 dB, which is more evident than removing the 3-D attention module. This phenomenon has represented the explored cross-scale nonlocal information that provides crucial information for SR tasks, which is suitable for RS image SR tasks.

V. Conclusion

This article proposes a RAN for RS image SR. RAN contains the degradation representation module based on the contrastive learning strategy. It avoids the problem of pixel-level degradation estimation modules relying on the accuracy of degradation. When there is an estimation deviation, the SR performance sharply declines. RS images are acquired in specific scenarios that have strong cross-scale self-similarity. This article designs a RAM containing a 3-D attention mechanism and a cross-scale feature aggregation module. The 3-D attention mechanism is based on channel and spatial attention mechanisms, which enhance feature interaction and explore local information. The cross-scale feature aggregation module introduces a graph convolutional neural network to improve the network's cross-scale feature extraction ability. Experimental results have shown that RAN handles SR tasks under different degradation conditions and reconstructs images with texture details. Furthermore, RS images cover various complex terrain environments. For future work, we will further consider the relief or terrain conditions for RS SR tasks.

Acknowledgment

Moreover, the authors sincerely appreciate editors, associated editors, and anonymous reviewers for their insightful suggestions and comments, which significantly improve this article.

References

[1] J. Wang, L. Du, Y. Li, G. Lyu, and B. Chen, "Attitude and size estimation of satellite targets based on ISAR image interpretation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5109015.

[2] D. Yu and S. Ji, "A new spatial-oriented object detection framework for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4407416.

[3] J. Wu, X. Su, Q. Yuan, H. Shen, and L. Zhang, "Multivehicle object tracking in satellite video enhanced by slow features and motion features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616426.

[4] Y. Cui, B. Hou, Q. Wu, B. Ren, S. Wang, and L. Jiao, "Remote sensing object tracking with deep reinforcement learning under occlusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605213.

[5] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Science Rev.*, vol. 232, Sep. 2022, Art. no. 104110.

[6] A. N. Shakya, I. K. Ramteke, and P. S. Wanjari, "Geo-spatial enabled water resource development plan for decentralized planning in India: Myths and facts," in *Mapping, Monitoring, and Modeling Land and Water Resources*. Boca Raton, FL, USA: CRC Press, 2021, pp. 179–196.

[7] S. Wang, T. Zhou, Y. Lu, and H. Di, "Contextual transformation network for lightweight remote-sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615313.

[8] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615611.

[9] L. Wang et al., "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10576–10585.

[10] Y.-S. Xu, S. R. Tseng, Y. Tseng, H.-K. Kuo, and Y.-M. Tsai, "Unified dynamic convolutional network for super-resolution with variational degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12493–12502.

[11] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3214–3223.

[12] S. A. Hussein, T. Tirer, and R. Giryes, "Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1425–1434.

[13] A. Shocher, N. Cohen, and M. Irani, "'Zero-shot' super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3118–3126.

[14] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3513–3522.

[15] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[16] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-D-Swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411415.

[17] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618014.

[18] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.

[19] J. Wang, Y. Zhong, and L. Zhang, "Change detection based on supervised contrastive learning for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601816.

[20] L. Bai, S. Du, X. Zhang, H. Wang, B. Liu, and S. Ouyang, "Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive learning and adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628313.

[21] J. Chen, D. Qin, D. Hou, J. Zhang, M. Deng, and G. Sun, "Multi-scale object contrastive learning-derived few-shot object detection in VHR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5635615.

[22] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 294–310.

[24] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[26] Z. Liu, R. Feng, L. Wang, W. Han, and T. Zeng, "Dual learning-based graph neural network for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628614.

[27] J. Chen, L. Jiao, X. Liu, L. Li, F. Liu, and S. Yang, "Automatic graph learning convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520716.

[28] B. Yang, F. Cao, and H. Ye, "A novel method for hyperspectral image classification: Deep network with adaptive graph structure integration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5523512.

[29] J. Bai et al., "Hyperspectral image classification based on superpixel feature subdivision and adaptive graph structure," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5524415.

[30] 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.

[31] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[32] L. Liebel and M. Körner, "Single-image super resolution for multi-spectral remote sensing data using convolutional neural networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 41, pp. 883–890, Jun. 2016.

[33] C. Tuna, G. Unal, and E. Sertel, "Single-frame super resolution of remote-sensing images by convolutional neural networks," *Int. J. Remote Sens.*, vol. 39, no. 8, pp. 2463–2479, Apr. 2018.

[34] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[35] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.

[36] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," *Remote Sens.*, vol. 10, no. 11, p. 1700, Oct. 2018.

[37] F. Deeba, F. A. Dharejo, Y. Zhou, A. Ghaffar, M. H. Memon, and S. Kun, "Single image super-resolution with application to remote-sensing image," in *Proc. Global Conf. Wireless Opt. Technol. (GCWOT)*, Oct. 2020, pp. 1–6.

[38] J. Zhang, S. Liu, Y. Peng, and J. Li, "Satellite image super-resolution based on progressive residual deep neural network," *J. Appl. Remote Sens.*, vol. 14, no. 3, Mar. 2020, Art. no. 032610.

[39] Z. Shao, L. Wang, Z. Wang, and J. Deng, "Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2663–2674, Aug. 2019.

[40] J. Gu, X. Sun, Y. Zhang, K. Fu, and L. Wang, "Deep residual squeeze and excitation network for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 15, p. 1817, Aug. 2019.

[41] X. Dong, Z. Xi, X. Sun, and L. Gao, "Transferred multi-perception attention networks for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 23, p. 2857, Dec. 2019.

[42] Y. Peng, X. Wang, J. Zhang, and S. Liu, "Pre-training of gated convolution neural network for remote sensing image super-resolution," *IET Image Process.*, vol. 15, no. 5, pp. 1179–1188, Apr. 2021.

[43] X. Dong, L. Wang, X. Sun, X. Jia, L. Gao, and B. Zhang, "Remote sensing image super-resolution using second-order multi-scale networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3473–3485, Apr. 2021.

[44] X. Li, D. Zhang, Z. Liang, D. Ouyang, and J. Shao, "Fused recurrent network via channel attention for remote sensing satellite image super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[45] S. Zhang, Q. Yuan, J. Li, J. Sun, and X. Zhang, "Scene-adaptive remote sensing image super-resolution using a multiscale attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4764–4779, Jul. 2020.

[46] Z.-X. Huang and C.-W. Jing, "Super-resolution reconstruction method of remote sensing image based on multi-feature fusion," *IEEE Access*, vol. 8, pp. 18764–18771, 2020.

[47] H. Wang, Q. Hu, C. Wu, J. Chi, and X. Yu, "Non-locally up-down convolutional attention network for remote sensing image super-resolution," *IEEE Access*, vol. 8, pp. 166304–166319, 2020.

[48] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[49] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[50] J. He, C. Dong, and Y. Qiao, "Interactive multi-dimension modulation with dynamic controllable residual learning for image restoration," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Nov. 2020, pp. 53–68.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[52] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 29–38.

[53] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.

[54] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, Jul. 2016, Art. no. 035004.

[55] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2010, pp. 270–279.

[56] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, Apr. 2012.

[57] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3262–3271.

[58] 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.

[59] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

Baodi Liu (Member, IEEE) received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2013.

He is currently an Associate Professor with the College of Control Science and Engineering, China University of Petroleum, Qingdao, China. His research interests include computer vision and machine learning.

Lifei Zhao received the bachelor's degree from the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China, in 2020, and the master's degree from the School of Ocean and Spatial Information, China University of Petroleum (East China), in 2023.

Her main research interests include computer vision and remote sensing image processing.

Shuai Shao (Member, IEEE) received the M.S. and Ph.D. degrees from the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China, in 2018 and 2022, respectively.

He was a Visiting Student at Tsinghua University, Beijing, China, from July 2019 to July 2020. Currently, he is a Post-Doctoral Researcher with the Zhejiang Laboratory, Hangzhou, China. During the Ph.D. degree, he published five articles as the first author in ACM Multimedia (ACMMM) and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His research interests include image processing, computer vision, and few-shot learning.

Weifeng Liu (Senior Member, IEEE) received the double B.S. degree in automation and business administration and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2002 and 2007, respectively.

He was a Visiting Scholar at the Center for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, Australia, from 2011 to 2012. He is currently a Professor with the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China. He has authored or coauthored a dozen papers in top journals and prestigious conferences, including ten ESI highly cited papers and three ESI hot papers. His research interests include pattern recognition and machine learning.

Dr. Liu is the Co-Chair of the IEEE SMC Technical Committee on Cognitive Computing. He is an Associate Editor of *Neural Processing Letters* and the Guest Editor of the Special Issue on "Signal Processing" by *IET Computer Vision*, *Neurocomputing*, and *Remote Sensing*. He also serves dozens of journals and conferences.

Dapeng Tao (Member, IEEE) is currently a Professor at the School of Information Science and Engineering, Yunnan University, Kunming, China, and with the Yunnan United Vision Technology Company Ltd., Kunming. He is mainly engaged in research in the field of artificial intelligence and has served as a Doctoral Advisor (computer science and technology) and a Doctoral Advisor (control science and engineering) at the University of the Chinese Academy of Sciences, Beijing, China.

Mr. Tao has served as a Special Reviewer and a Guest Editor for more than ten international academic journals, including IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING.

Weijia Cao received the master's and Ph.D. degrees in computer science from the University of Macau, Macau, China, in 2013 and 2017, respectively.

She is currently an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her main research interests revolve around machine learning and remote-sensing image processing.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees from Tufts University, Medford, MA, USA, in 2008 and 2010, respectively, all in electrical engineering.

He is currently a Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the International Society for Optical Engineering (SPIE) and a Senior Member of the China Computer Federation (CCF). He was a recipient of the Third Price of Macao Natural Science Award in 2014 and 2020. He is a Co-Chair of the Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He was listed as World's Top 2% Scientists on the Stanford University Releases List in 2020 and 2021 and the Highly Cited Researcher in the Web of Science in 2020 and 2021. He is an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and four other journals.