From Artifact Removal to Super-Resolution

Jiaming Wang, Zhenfeng Shao, Xiao Huang, Tao Lu, *Member, IEEE*, Ruiqian Zhang, and Yong Li

***Abstract*— Deep-learning-based super-resolution (SR) methods have been extensively studied and have achieved significant performance with deep convolutional neural networks. However, the results still suffer from the ringing effect, especially in satellite image SR tasks, due to the loss of image details in the satellite degradation process. In this article, we build a novel satellite SR framework by decomposing a high-resolution image into three components, i.e., low-resolution (LR), artifact, and high-frequency information. Specifically, we propose an artifact removal network with a self-adaption difference convolution (SDC) to fully exploit the structure prior in the LR image and predict the artifact map. Considering that the artifact map and the high-frequency map share a similar pattern, we introduce the supervised structure correction (SSC) block that establishes a bridge between the high-frequency generation process and the artifact removal process. Experimental results on satellite images demonstrate that the proposed method owns an improved tradeoff between the performance and the computational cost compared to existing state-of-the-art satellite and natural SR methods. The source code is available at https://github.com/jiaming-wang/ARSRN.**

***Index Terms*— Artifact removal, difference convolution, remote sensing, super-resolution (SR).**

I. Introduction

As an emerging means of Earth observation, satellite platforms have drawn increasing attention in both military and civil fields, such as surface feature segmentation [1], [2], environmental monitoring [3], [4], satellite mapping [5], and object detection [6], [7]. Limited by the under-sampling effect of imaging sensors and complexity in the degradation process, captured satellite images (especially those from geostationary satellites) may fail to meet the demand of many applications that require high-precision measurement. Image super-resolution (SR) is a software-level technology that aims to increase the spatial resolution of low-resolution (LR) images without introducing additional hardware costs. With years of development, satellite image SR has become an essential step for many real-time and high-precision applications.

Many deep-learning-based SR methods have been proposed to learn the nonlinear mapping function, usually trained with LR images and the corresponding high-resolution (HR) images [8]. The residual learning strategy [9] was introduced into the SR task to learn the missing high-frequency information, which can greatly reduce the complexity of optimization. To increase the overall sharpening effect, some studies introduced edge-preserving filters for detail enhancement by specifically considering image edges [10], [11], [12]. Yang *et al.* [10] introduced an edge-guided recurrent SR framework to model edge priors and predict edges of the HR images. Mao *et al.* [11] proposed a generative adversarial network (GAN) [13] with an edge-preserving strategy to preserve the edge structures and reconstruct visually pleasing textures. Jiang *et al.* [12] proposed an edge-enhancement GAN that retains details with removed noises. Compared with natural images, the detail losses of the acquired satellite are severer due to the process of image degradation [12]. The above-mentioned edge-preserving algorithms assume that the LR edge priors can be used for inferring the missing high-frequency information. However, given the large information gap between LR and HR images, existing edge-preserving methods usually fail to generate sheer textures from the artifact-contaminated edge, leading to spectral distortion and artifact. Therefore, it remains to be a challenge to develop efficient models with texture restoration capability without being affected by the artifacts.

As shown in Fig. 1, the LR image suffers from notable serrated and blurry artifacts in the contour part influenced by the ringing effect (see the edges of the aircraft and the road line). With the increase of the sampling factor, the ringing effect becomes more notable, which pollutes the high-frequency information of the image. Therefore, we attempt to decompose the HR image into LR, artifact, and high-frequency images. For an intuitive visual contrast, the ImageEnhance module is employed to enhance the contrast of the sparse artifact and high-frequency images. The artifact image is defined as the difference map (i.e., $\text{relu}(LR - HR)$) between the LR and HR images. Meanwhile, the high-frequency image is defined as the difference map (i.e., $\text{relu}(HR - LR)$) between the HR and LR images. We observe that the artifact map and the high-frequency map are similar in structure and complementary in detail, as shown in Fig. 1. Specifically, the high-frequency image describes objects' edges, while the artifact image highlights the fringes of the high-frequency image. Therefore, we hypothesize that the artifact image

Manuscript received 16 April 2022; revised 21 June 2022 and 18 July 2022; accepted 2 August 2022. Date of publication 5 August 2022; date of current version 16 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42090012; in part by the Guangxi Science and Technology Program under Grant GuiKe 2021AB30019; in part by 03 Special Research and 5G Project of Jiangxi Province in China under Grant 20212ABC03A09; in part by the Zhuhai Industry University Research Cooperation Project of China under Grant ZH22017001210098PWC; in part by the Sichuan Science and Technology Program under Grant 2022YFN0031; and in part by the Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence under Grant ZZJJ202202. *(Corresponding author: Zhenfeng Shao.)*

Jiaming Wang, Zhenfeng Shao, and Yong Li are with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: wjmecho@whu.edu.cn; shaozhenfeng@whu.edu.cn; yong.li@whu.edu.cn).

Xiao Huang is with the Department of Geosciences, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: xh010@uark.edu).

Tao Lu is with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China (e-mail: lutxyl@gmail.com).

Ruiqian Zhang is with the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing 100036, China (e-mail: zhangruiqian@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3196709

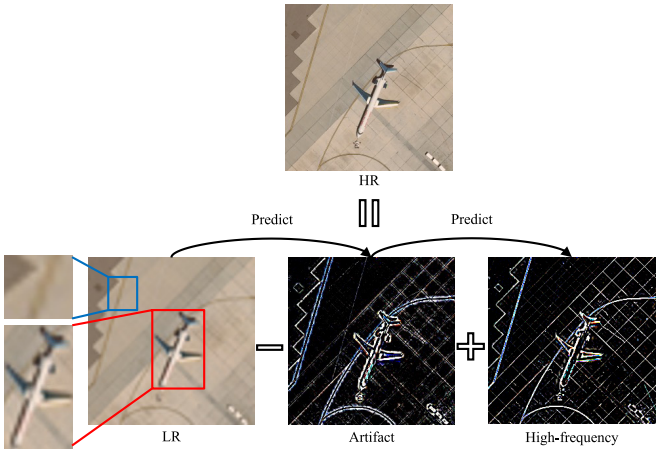


Fig. 1. Proposed SR strategy. We decompose information from a HR image into LR information, artifact information, and high-frequency information. Note that the artifact map and the high-frequency map are similar in structure and complementary in detail.

can be regarded as the prior information that facilitates the prediction of high-frequency information.

According to the above observation, we propose a refined image decomposition prior model, in which HR images are decomposed to LR, artifact, and high-frequency images, respectively. According to this concept, the proposed framework, named artifact removal to SR network (ARSRN), includes an artifact removal network and a high-frequency generation network, as shown in Fig. 2. Different from existing methods, the proposed method decouples the SR task into two more refined processes, i.e., artifact removal and high-frequency generation, thus fully exploring the image as prior information more fully and generating better-reconstructed results. Specifically, the proposed strategy extracts the artifact of the LR satellite image as the prior information used to guide the generation of high-frequency information. These two subnetworks are hourglass-shaped with proposed self-adaption difference convolutions (SDCs). The predicted results of these two subnetworks are artifact and high-frequency images, respectively. In addition, we design a supervised structure correction (SSC) block that establishes a bridge between the artifact image and the high-frequency image, given the fact that these two images share a similar image structure. In addition, we introduce a fast Fourier loss function (FFL) to calibrate the difference between reconstructed images and ground-truthing images in the frequency domain. Comprehensive ablation studies demonstrate the effectiveness of each subnetwork in the proposed ARSRN. Experiments that compare the proposed ARSRN with state-of-the-art SR methods prove that ARSRN achieves a satisfactory tradeoff between performance and efficiency.

The main contributions of the proposed ARSRN are:

1) We propose an efficient SR framework that divides the SR process into two tasks, i.e., artifact removal and high-frequency generation. Both qualitative and quantitative results demonstrate the superiority of the proposed method over state-of-the-art algorithms. Besides the satellite SR task, we also conduct experiments on a natural dataset to better reflect the effectiveness and expandability of the proposed ARSRN framework.
2) We design an SDC to capture the artifact and high-frequency information. The proposed difference convolution is effective in describing the fine-grained texture information, thus significantly reducing the ringing effect during the image degradation process.
3) We propose a SSC block to forge links between the artifact removal task and the high-frequency generation task in the proposed framework. The proposed SSC block regards artifact as the prior information that assists the prediction of high-frequency information.

The remaining part of this article is organized as follows. Section II presents the related work of image SR and the difference convolution neural network (DCNN). Section III details the proposed ARSRN. Section IV reports experimental results. Section V concludes this study.

II. Related Work

In this section, we briefly review the related works on satellite image SR methods and DCNNs.

A. Satellite Image SR

Image SR has been greatly advanced in the last couple of years. Dong *et al.* [14] introduced the convolutional neural networks (CNNs) into the SR task by constructing a three-layers nonlinear CNN, i.e., SRCNN. Furthermore, Liebel and Körner [15] employed SRCNN for the high radiation resolution and multispectral resolution satellite image SR using 13-band satellite images captured from the Sentinel-2 satellite. To better utilize local information often ignored by deep networks, Lei *et al.* [16] proposed a local–global combined network that learns multilevel representations of satellite images and fuses local details and global environmental priors. Inspired by GoogLeNet [17] and Qin *et al.* [18] introduced a CNN with multiscale kernels to describe the multiangle features. Enlightened by the immense success of residual learning [9] in a series of computer vision tasks [19], [20], [21], Lu *et al.* [22] found that multiscale residual neural networks are able to reconstruct the desired fine edge/detail textures. Regarding the vanishing-gradient problem and the inefficient feature reuse strategy, Ma *et al.* [23] achieved improved performance by proposing a residual dense back-projection network that structures local and global residual with a large sampling factor. Similarly, Dong *et al.* [24] designed a dense sampling mechanism with the channel attention mechanism to convolve prior knowledge from different depths. Haut *et al.* [25] introduced residual learning into channel attention to encourage the model to learn high-frequency information and, at the same time, suppress low-frequency information. Zhang *et al.* [26] proposed high-order attention to learn the attention map of different convolution layers with refined fused features. The unique characteristics of remote sensing images (e.g., different bodies with the same spectrum and diversity in object scales) make the SR task challenging. Zhang *et al.* [27] performed model fine-tuning for different scenes based on transfer learning and allocated models for SR tasks in different scenes in an adaptive manner. Lei and Liu [28] combined the inception block [17] and spatial-channel

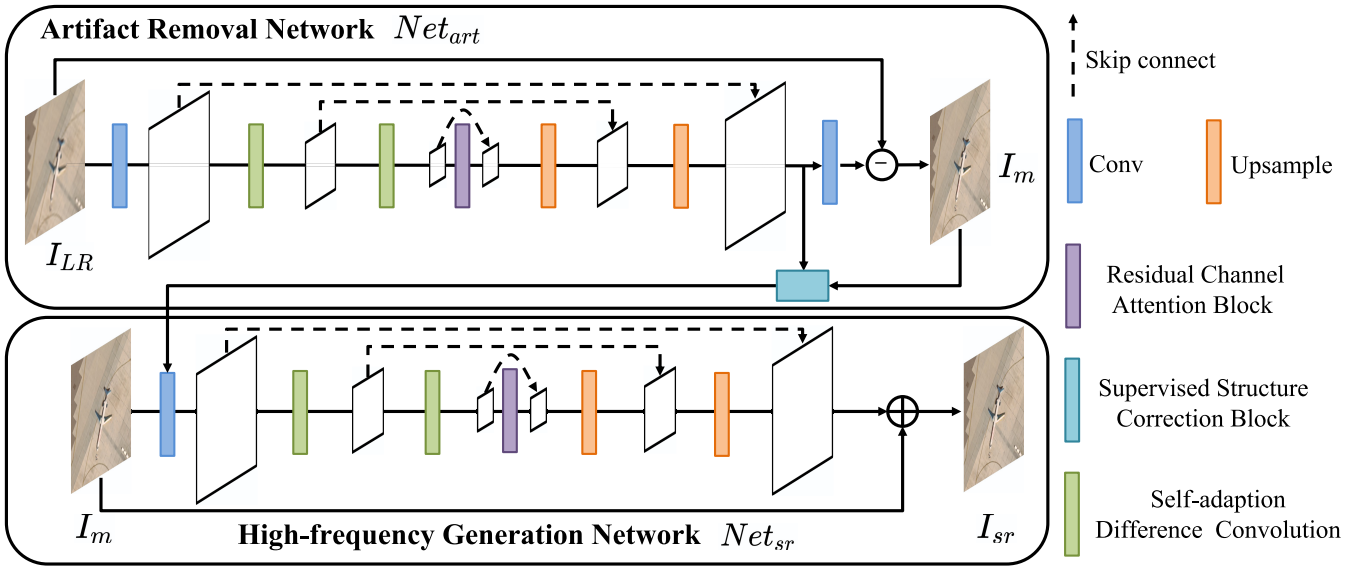


Fig. 2. Proposed SR framework includes an artifact removal network and a high-frequency generation network, while the SSC block builds a bridge between these two networks. The artifact removal network can predict the artifact information I_{art} of the LR image I_{LR}. I_m ($I_m = I_{LR} - I_{\text{art}}$) represents the deartifact satellite image. The high-frequency generation network can predict the high-frequency information I_{fq} of the deartifact image I_m. I_{sr} ($I_{sr} = I_m + I_{fq}$) represents the reconstruction image. $\oplus$ and $\ominus$ denote element-wise addition and element-wise subtraction, respectively.

attention module to screen the important spatial-channel features.

Despite the great subjective and objective performances of the above CNN-based algorithms, the issue of insufficient edge texture still remains. To mitigate this problem, Ledig *et al.* [29] designed a GAN framework [13] to approximate the probability distribution of the ground truth over the manifold space. However, GAN-based methods designed for the SR task tend to be sensitive to noises, leading to the production of irrelevant high-frequency information. Jiang *et al.* [12] proposed an edge-enhancement GAN for satellite image SR to alleviate edge blur in the reconstructed SR images. Lei *et al.* [30] proposed a coupled-discriminate GAN with a shuffle gate function to enhance the identification of low-frequency information.

In multispectral and hyperspectral SR tasks, fusion-based methods are the mainstream research direction. Masi *et al.* [31] first proposed CNN to learn the complex mapping function between the LR multispectral and the HR multispectral image. Qu *et al.* [32] proposed an unsupervised encoder-decoder algorithm that encourages the network to explore information with sparse Dirichlet distribution. Xu *et al.* [33] proposed a surface- and deep-level constraint to reconstruct the spatial and spectral information and effectively avoided information distortion. Wang *et al.* [34] proposed a dual-path network to capture global textural and spectral information by integrating the neighbor embedding strategy. Qu *et al.* [35] proposed an unregistered and unsupervised mutual framework to fuse hyperspectral images and multispectral images without multimodality registration.

B. Difference Convolution Neural Networks

He *et al.* [9] proposed the residual block with shortcut connections, serving as an equivalent to a shallow network. Another alternative is to build the block guided by the finite difference strategy, serving as an approximate for the difference equation [36]. Then, Jing *et al.* [37] proposed a succinct DCNN unit with difference mapping to learn the difference salient features. Furthermore, more variants were proposed considering directional changes among a pixel and its neighbors [38]. The DCNN is able to analyze the pattern orientations of a pixel and its neighbors by applying predefined constant filters (e.g., Roberts, Sober, and LoG operators) used as the edge extractors. Inspired by the local binary patterns (LBP), Yu *et al.* [39] proposed a central difference convolution (CDC) that calculates the difference map as the texture information between the central pixel and the other eight neighboring pixels. To minimize feature redundancy, Yu *et al.* [40] decoupled the CDC into horizontal-vertical and diagonal uncrossed suboperators and proposed a cross-CDC to increase computational efficiency. Inspired by the extended-LBP [41], Su *et al.* [42] employed the parameters in the predefined constant filters as the initial CNN parameters. Other efforts [43], [44] have been made to combine graph convolution and CDC and apply the combined model to numerous computer vision tasks [45], [46], [47].

III. Proposed Method

In this section, we provide details for the problem formulation, the artifact removal network, the high-frequency generation network, and the loss function.

A. Problem Formulation

Fig. 2 presents the network architecture of the proposed method. In general, the proposed framework consists of an artifact removal network and a high-frequency generation network. Specifically, given a LR image $I_{lr} \in \Re^{C\times H\times W}$ and the corresponding ground truth (the original HR version image) $I_{HR} \in \Re^{C\times tH\times tW}$, the goal of the proposed framework is to generate the SR image I_{SR}. Meanwhile, the bicubic version of

a LR image is defined as $I_{LR} \in \Re^{C\times tH\times tW}$. H, W, C indicate the height, the width, and the channel of I_{lr}, respectively. t suggests the scale factor. The goal of the proposed model is to reconstruct the HR image

$$I_{SR} = \mathcal{F}(I_{LR}) \tag{1}$$

where $\mathcal{F}(.)$ denotes the SR function of the proposed end-to-end framework.

Different from existing SR methods that treat the SR task as a mapping function between the LR and HR images/residuals, we divide the process into two refined procedures, i.e., artifact removal and high-frequency generation. Such a design allows us to exploit the structure information in the artifact image so that the edge prediction can be guided. The proposed subnetworks are equipped with the designed SDC layers and connected via the proposed SSC block. In order to alleviate the burden of the proposed method, we adopt the hourglass-shaped network for multiscale reconstruction. We provide details for these two subnetworks in the subsections.

1) Artifact Removal Network: First, the bicubic version of the LR image, I_{LR}, is fed into an hourglass-shaped network equipped with the SDC layer. This process can be described as

$$\mathcal{M}_0^{(a)} = \mathcal{F}_{f1}(I_{LR}) \tag{2}$$

where $\mathcal{F}_{f1^{(a)}}(.)$ denotes 3×3 convolution operation as the initial feature extraction layer. Then, $\mathcal{M}_0^{(a)}$ is fed into the proposed SDC

$$\mathcal{M}_1^{(a)} = \mathcal{F}_{sdc_1^{(a)}}\left(\mathcal{M}_0^{(a)}\right) \tag{3}$$

$$\mathcal{M}_2^{(a)} = \mathcal{F}_{\mathrm{sdc}_2^{(a)}}\left(\mathcal{M}_1^{(a)}\right) \tag{4}$$

where $\mathcal{F}_{\mathrm{sdc}}(.)$ denotes the function of the proposed SDC with kernel size 3×3 and stride 2, serving as a downsampling layer. Particularly, we add a nonlinear mapping module in the middle of the network

$$\mathcal{M}_3^{(a)} = \mathcal{F}_{ft}^{(a)}\left(\mathcal{M}_2^{(a)}\right) \tag{5}$$

where $\mathcal{F}_{ft}(.)$ denotes the feature transformation module with R_n representing residual channel attention blocks [48]. Then, the multiscale features are restored to the original resolution, as

$$\mathcal{M}_4^{(a)} = \mathcal{F}_{up_1^{(a)}}\left(\mathrm{concat}\left(\mathcal{M}_2^{(a)}, \mathcal{M}_3^{(a)}\right)\right) \tag{6}$$

$$\mathcal{M}_5^{(a)} = \mathcal{F}_{up_2^{(a)}}\left(\mathrm{concat}\left(\mathcal{M}_1^{(a)}, \mathcal{M}_4^{(a)}\right)\right) \tag{7}$$

where $\mathcal{F}_{up}(.)$ denotes the PixelShuffle operator [49] and concat(.) refers to the concatenation operation. We employ a reconstruction layer for artifact reconstruction

$$I_{\mathrm{art}} = \mathcal{F}_{\mathrm{rec}}^{(a)}\left(\mathcal{M}_5^{(a)}\right) \tag{8}$$

where I_{art} represents the artifact image. The output image can be expressed as

$$I_m = I_{LR} - I_{\mathrm{art}} \tag{9}$$

where I_m represents the deartifact satellite image.

The artifact removal image and final feature map are fed into the SSC block, which can be formulated as

$$M_{\mathrm{ssc}} = \mathcal{F}_{\mathrm{ssc}}\left(I_m, \mathcal{M}_5^{(a)}\right) \tag{10}$$

where $\mathcal{M}_{\mathrm{ssc}}$ represents the feature map for SSC in the high-frequency domain (more details in Section III-C).

2) High-Frequency Generation Network: After the artifact removal, we feed $\mathcal{M}_{\mathrm{ssc}}$ and I_m into the high-frequency generation network, which is similar to the artifact removal network. First, we employ the $\mathcal{M}_{\mathrm{ssc}}$ to direct the network's attention to the region of interest in I_{art}, as image edge prior. In particular, this process can be formulated as

$$\mathcal{M}_0^{(h)} = \mathrm{concat}\left(\mathcal{F}_{f1^{(h)}}(I_m), \mathcal{M}_{\mathrm{ssc}}\right) \tag{11}$$

where $\mathcal{F}_{f1^{(h)}}(.)$ denotes 3×3 convolution operation as the initial feature extraction layer. Then, the high-frequency image I_{fq} is generated in the same process as the artifact removal network

$$\mathcal{M}_1^{(h)} = \mathcal{F}_{\mathrm{sdc}_1^{(h)}}\left(\mathcal{M}_0^{(h)}\right) \tag{12}$$

$$\mathcal{M}_2^{(h)} = \mathcal{F}_{\mathrm{sdc}_2^{(h)}}\left(\mathcal{M}_1^{(h)}\right) \tag{13}$$

$$\mathcal{M}_3^{(h)} = \mathcal{F}_{ft}^{(h)}\left(\mathcal{M}_2^{(h)}\right) \tag{14}$$

$$\mathcal{M}_4^{(h)} = \mathcal{F}_{up_1^{(h)}}\left(\mathrm{concat}\left(\mathcal{M}_2^{(h)}, \mathcal{M}_3^{(h)}\right)\right) \tag{15}$$

$$\mathcal{M}_5^{(h)} = \mathcal{F}_{up_2^{(h)}}\left(\mathrm{concat}\left(\mathcal{M}_1^{(h)}, \mathcal{M}_4^{(h)}\right)\right) \tag{16}$$

$$I_{fq} = \mathcal{F}_{\mathrm{rec}}^{(f)}\left(\mathcal{M}_5^{(h)}\right). \tag{17}$$

Finally, the SR image I_{SR} can be formulated as

$$I_{SR} = I_m + I_{fq}. \tag{18}$$

B. SDC Block

In the image degradation process, the ringing effect tends to appear on regions with rapidly changed gray scales in the contaminated image. The ringing effect is more notable in LR satellite images, as single pixels in LR images are more distant from each other compared to the ones in HR images. Existing manually designed deep-learning-based networks tend to ignore the prior information in the ringing effect. The proposed SDC is shown in Fig. 3. The artifact removal network can employ the proposed SDC to predict artifact image, which can guide the high-frequency image generation by the SDC in the high-frequency generation network.

The SDC considers the incoming features $\mathcal{M}_{in}$ in the generation of the texture image $\mathcal{M}_{\mathrm{out}}$

$$\mathcal{M}_{\mathrm{out}} = \mathcal{F}_{\mathrm{sdc}}(\mathcal{M}_{in}) \tag{19}$$

where $\mathcal{F}_{\mathrm{sdc}}(.)$ refers to the function of the SDC.

Given the existence of redundant information in the differential operation of the neighborhood features, we design the self-adaption strategy as show in the rectangle of Fig. 3. Specifically, the proposed strategy can adaptively adjust the size of the difference operator, such as retaining the top k_n largest coefficients, to reduce the amount of calculation and parameters. In the local receptive field region $\mathcal{R}$, the output

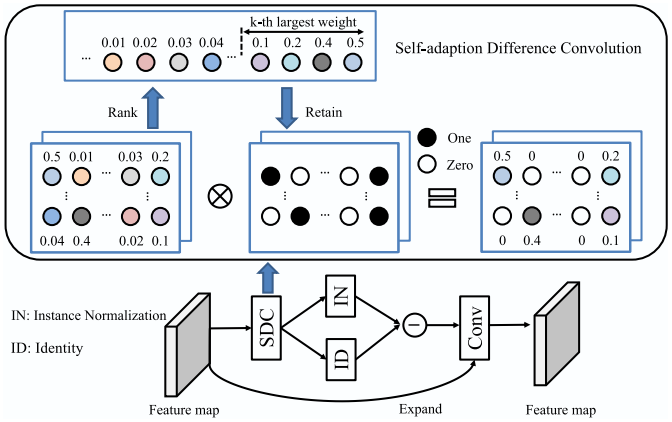


Fig. 3. Proposed SDC.

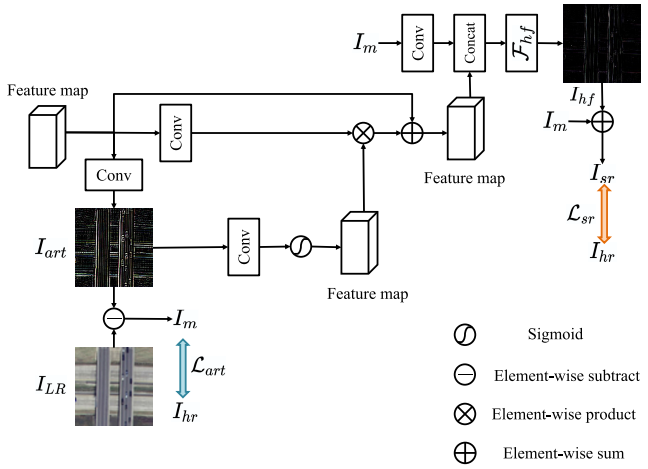


Fig. 4. Proposed SSC block.

feature map y of the SDC can be generated from the input map $\mathcal{M}_{in}$ and can be formulated as

$$\tilde{w} = \begin{cases} w_{id}, & id \in \text{topK}\{w\} \\ 0, & \text{other} \end{cases} \tag{20}$$

$$y(p_0) = \sum_{p_n \in \mathcal{R}} \tilde{w}(p_0)(\mathcal{M}_{in}(p_n + p_0) - \mathcal{M}_{in}(p_0)) \tag{21}$$

where p_0 denotes the location information and p_n enumerates the locations in the region $\mathcal{R}$. w denotes the original weight map and $\tilde{w}$ denotes the adaptively adjusted weight map. topK$\{w\}$ denotes the top k_n largest numbers of w with number index id.

Then, the feature map y is divided into two feature maps with the same shape, i.e., $\mathcal{M}_{\text{mid1}}, \mathcal{M}_{\text{mid2}}$

$$[\mathcal{M}_{\text{mid1}}, \mathcal{M}_{\text{mid2}}] = y. \tag{22}$$

The feature map $\mathcal{M}_{\text{mid1}}$ is fed into the instance normalization (IN) that normalizes the mean and variance of input features to learn affine parameters. Then, the recalibrated $\mathcal{M}_{\text{mid1}}$ is concatenated with $\mathcal{M}_{\text{mid2}}$ as

$$\mathcal{M}_{\text{mid}} = \text{concat}(\mathcal{F}_{IN}(\mathcal{M}_{\text{mid1}}), \mathcal{M}_{\text{mid2}}) \tag{23}$$

where $\mathcal{F}_{IN}(.)$ denotes the function of the IN layer that remains independent between image instances.

Finally, the output can be formulated as

$$\mathcal{M}_{\text{out}} = \mathcal{M}_{\text{mid}} - \theta \mathcal{F}_{\text{conv}}(\mathcal{M}_{in}) \tag{24}$$

where θ is the control parameter ($\theta = 0.8$).

C. SSC Block

Typically, the goal of SR is to predict high-frequency information and generate pleasant visual effects. Considering the complementary nature between the artifact image and the residual image, we introduce a SSC block, as shown in Fig. 4, to connect these two networks. Different from the series connection strategy in the existing two-stage framework, the proposed SSC block serve as a direct bridge between the artifact removal network and the high-frequency network, aiming to utilize the artifact prior information. The SSC strategy is mainly composed of two parts: 1) structure information transmission and 2) supervised modification. The description of the supervised modification is detailed in the loss function section.

In the structure information transmission phase, the dearti-fact image (I_m) is fed into an attention block to capture the structure information of the artifact. Then, the attention feature map is fed into the high-frequency generation network to guide texture generation. Specifically, the output artifact image is fed into the convolutional and the sigmoid layers, which can be formulated as

$$M_{\text{att}} = \mathcal{F}_{\text{sig}}(\mathcal{F}_{\text{att}}(I_{art})) \tag{25}$$

where $\mathcal{F}_{\text{att}}(.)$ denotes a 3×3 convolution operation and $\mathcal{F}_{\text{sig}}(.)$ denotes the sigmoid operation that normalizes the features. M_{att} is the attention map of the artifact image. Then, the SSC feature maps M_{ssc} can be formulated as

$$M_{\text{ssc}} = M_{\text{att}} \times \mathcal{M}_5^{(a)} + \mathcal{M}_5^{(a)}. \tag{26}$$

D. Loss Function

In the training phase of the artifact removal network, the loss function is defined by the mean absolute error function between the artifact images and the residual images

$$\mathcal{L}_{\text{art}}(\Theta_1) = \frac{1}{N} \sum_{n=1}^{N} \| I_{\text{art}}^n - \mathcal{G}\left(I_{LR}^n - I_{hr}^n\right) \|_1 \tag{27}$$

$$\mathcal{G}(x) = \begin{cases} x, & \text{if} \quad x > 0 \\ 0, & \text{if} \quad x < 0 \end{cases} \tag{28}$$

where I_{art}^n, I_{LR}^n, and I_{HR}^n are the nth artifact image, LR image, and ground truth satellite image, respectively. N denotes the number of images in the batch, and Θ_1 refers to the parameter set of the artifact removal network.

In recent literature, pixel domain-oriented loss functions, such as mean average error (MAE) and mean squared error (MSE) are the most commonly used loss functions in the SR task. The goal of the image SR task is to generate missing high-frequency information. In addition, the Fourier domain provides a better description of the global information of the image [55]. In order to simultaneously ensure the frequency credibility of the SR results, the high-frequency generation

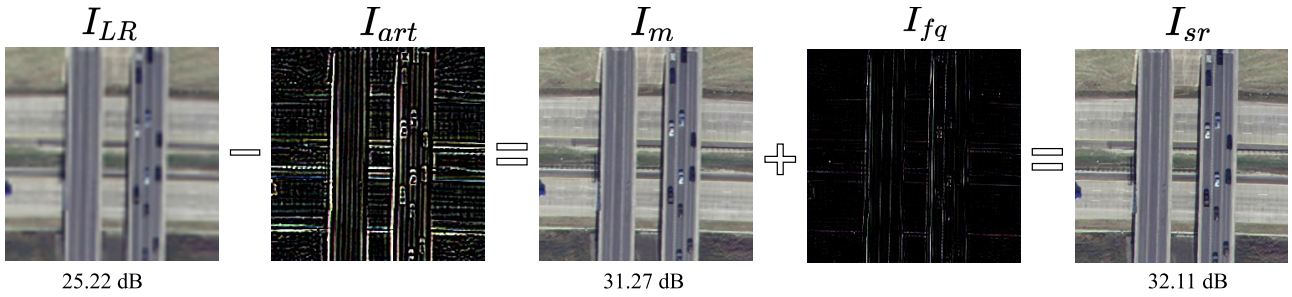


Fig. 5. Visual representation of the proposed method.

network is optimized via SR loss function and FFL. These functions are defined as

$$\mathcal{L}_{sr}(\Theta_2) = \frac{1}{N}\sum_{n=1}^{N}\left\|I_{sr}^{n} - I_{hr}^{n}\right\|_1 \tag{29}$$

$$\mathcal{L}_{\text{ffl}}(\Theta_2) = \frac{1}{N}\sum_{n=1}^{N}\left\|\varphi\left(I_{sr}^{n}\right) - \varphi\left(I_{hr}^{n}\right)\right\|_1 \tag{30}$$

where the SR loss function $\mathcal{L}_{sr}$ denotes the similarity between reconstructed images and the ground-truthing images, while the FFL loss function $\mathcal{L}_{\text{ffl}}$ denotes the similarity in the frequency domain. I_{sr}^{n} is the nth reconstructed image. $\varphi(.)$ denotes the FFL and Θ_2 refers to the parameter set of the high-frequency generation network.

The final loss function is a weighted sum of the artifact loss $\mathcal{L}_{\text{art}}$, the high-frequency generation loss $\mathcal{L}_{sr}$, and the FFL loss $\mathcal{L}_{\text{ffl}}$ as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{art}} + \mathcal{L}_{sr} + \lambda_{\text{ffl}}\mathcal{L}_{\text{ffl}} \tag{31}$$

where the hyperparameter λ_{ffl} is used to balance the contributions of different loss terms.

To provide a better understanding of the proposed ARSRN, we present selected intermediate results in Fig. 5. The results suggest that the reconstructed image I_{sr} presents sharper edges with reconstructed visually pleasant details, thanks to the intervention of high-frequency information. Even without additional high-frequency information, the reconstructed image I_m presents satisfactory details.

IV. Experiments

A. Datasets and Implementation Details

We compare the performance of the proposed method with the ones from recent deep-learning-based SR methods (residual dense network (RDN) [52], deep back-projection network (DBPN) [53], residual channel attention network (RCAN) [48], and second-order attention network (SAN) [54]) and recent deep-learning-based satellite SR methods (mixed high-order attention network (MHAN) [26]) on three satellite SR datasets, i.e., University of California (UC) Merced Land Use Dataset (UC Merced)[1] [56], remote sensing scene classification (RSCNN7)[2] [57], and University of Chinese Academy of Sciences (UCAS)-high resolution aerial object detection dataset (UCAS-AOD).[3] Five widely used image quality assessment (IQA) metrics (i.e., peak signal-to-noise ratio (PSNR), structural similarity (SSIM), feature similarity (FSIM), learned perceptual image patch similarity (LPIPS) [50], and mean perceptual score (MPS) [51]) are employed to quantify model performances. The proposed model is trained on the desktop with Ubuntu 18.04, compute unified device architecture (CUDA) 10.2, CUDA deep neural network (CUDNN) 7.5, and three Nvidia TITAN graphics processing units (GPUs). These two subnetworks are trained with the same parameters. The Adam optimizer is employed for optimization with a mini-batch size of 16, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e-8$. The learning rate is initialized to $1e-4$ and decreases by a factor of 10 for every 500 epochs. The original images are downsampled with the bicubic function.

[1]http://weegee.vision.ucmerced.edu/datasets/landuse.html
[2]https://github.com/palewithout/RSSCN7
[3]https://github.com/bingoxumo/UCAS-AOD_FFDP

B. Comparing With State-of-the-Arts on UC Merced Dataset

The UC Merced dataset is a popular dataset for remote sensing image classification. The UC Merced dataset contains crowdsourced images of 21 categories with a benchmark of 256×256 pixels from massive crowdsource data. The spatial resolution of the UC Merced dataset is 0.3 m. We randomly select 90 images and 10 images from each category for training and testing, respectively.

Table I reports the average performance of the proposed method and other competing deep-learning-based algorithms on the UC Merced dataset with scale factors $\times 2$ and $\times 4$, where the bold values denote the best results and underlined values represent the second-best results. The average PSNR values of the proposed method are 0.09/0.21 dB higher than that of the second-best method with upsampling factors $t = 4$ and $t = 8$, respectively. The average MPS reflect the notable improvement of the proposed method, both at the pixel level and the perception level. As the sampling factor increases, the artifacts caused by the degradation and upsampling processes become notable. The proposed method is able to suppress the generation of artifacts, leading to its superior performance under scenarios of high sampling factors. One limitation of the proposed method is that its computational cost tends to considerably increase with the increase of the upsampling factor t.

Figs. 6 and 7 show the reconstructed images on the UC Merced dataset from different comparison algorithms with upsampling factors $\times 2$ and $\times 4$, respectively. Images in the last row show the MSE between the reconstructed images and the ground-truthing images. We observe that the proposed method outperforms other competing algorithms, evidenced by

TABLE I

QUANTITATIVE COMPARISON OF SEVEN METHODS ON THE UC MERCED DATASET. BEST AND SECOND-BEST SCORES ARE **HIGHLIGHTED** AND UNDERLINED, RESPECTIVELY. ↑ INDICATES THAT THE LARGER THE VALUE, THE BETTER THE PERFORMANCE, AND ↓ INDICATES THAT THE SMALLER THE VALUE, THE BETTER THE PERFORMANCE

Methods	Scales	#Params/M	FLOPs/G	PSNR ↑	SSIM↑	FSIM↑	LPIPS[50] ↓	MPS [51]↑
Bicubic	× 2	-	-	31.67	0.8806	0.9156	0.1928	0.8439
RDN [52]		1.00	2.32	34.44	0.9287	0.9582	0.1016	0.9136
DBPN [53]		5.95	34.76	35.01	0.9347	0.9629	0.0840	0.9254
RCAN [48]		15.44	35.36	<u>35.07</u>	<u>0.9359</u>	<u>0.9632</u>	**0.0754**	**0.9302**
SAN [54]		15.67	36.10	35.05	0.9344	0.9632	0.0754	0.9295
MHAN [26]		11.20	26.10	34.97	0.9342	0.9626	0.0868	0.9237
Our		2.75	7.93	**35.16**	**0.9360**	**0.9643**	<u>0.0756</u>	<u>0.9302</u>
Bicubic	× 4	-	-	26.88	0.6871	0.7663	0.4781	0.6045
RDN [52]		1.15	3.73	28.48	0.7588	0.8342	0.3211	0.7189
DBPN [53]		10.43	92.68	<u>28.58</u>	<u>0.7702</u>	**0.8459**	0.2732	0.7428
RCAN [48]		15.59	36.77	28.36	0.7614	0.8435	0.2692	<u>0.7505</u>
SAN [54]		15.82	37.51	28.39	0.7613	0.8431	<u>0.2629</u>	0.7493
MHAN [26]		11.35	28.53	28.42	0.7626	0.8383	0.2879	0.7374
Our		2.75	31.74	**28.79**	**0.7719**	<u>0.8457</u>	**0.2627**	**0.7546**

Parameters (Params) and floating-point operations per second (FLOP) are tested on an LR image with 48 × 48 pixels.

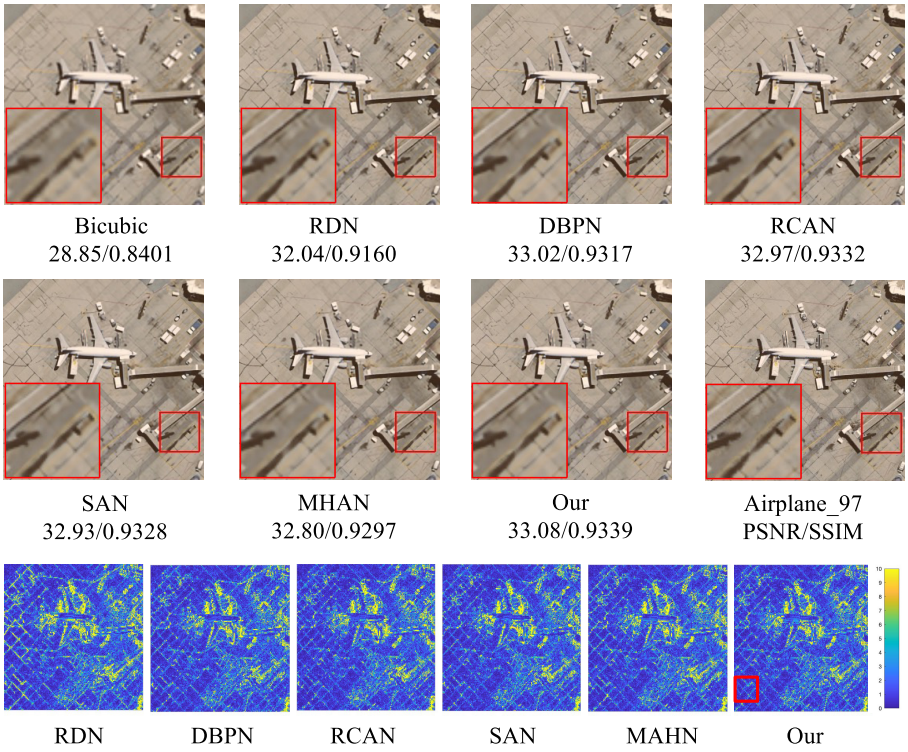


Fig. 6. Qualitative comparison of the proposed method with six counterparts on a typical satellite image pair from the UC Merced dataset with an upsampling factor of ×2. Images in the last row visualize the MSE between the SR results and the ground truth. (Zoomed-in view to see more details.)

its better recovery of textures and structures, e.g., the outline of vehicles and roadside lines. Specifically, satellite images with complex textures and dense objects (i.e., the parking lot with a large number of vehicles) can be easily polluted by artifacts. The artifact removal operation in the proposed model leads to accurate contours, as demonstrated in the MSE map.

C. Comparing With State-of-the-Arts on RSCNN7 Dataset

The RSCNN7 is another popular dataset for remote sensing image classification. It contains images (400 × 400 pixels) of seven typical scene categories collected from Google Earth. We randomly select 300 images and

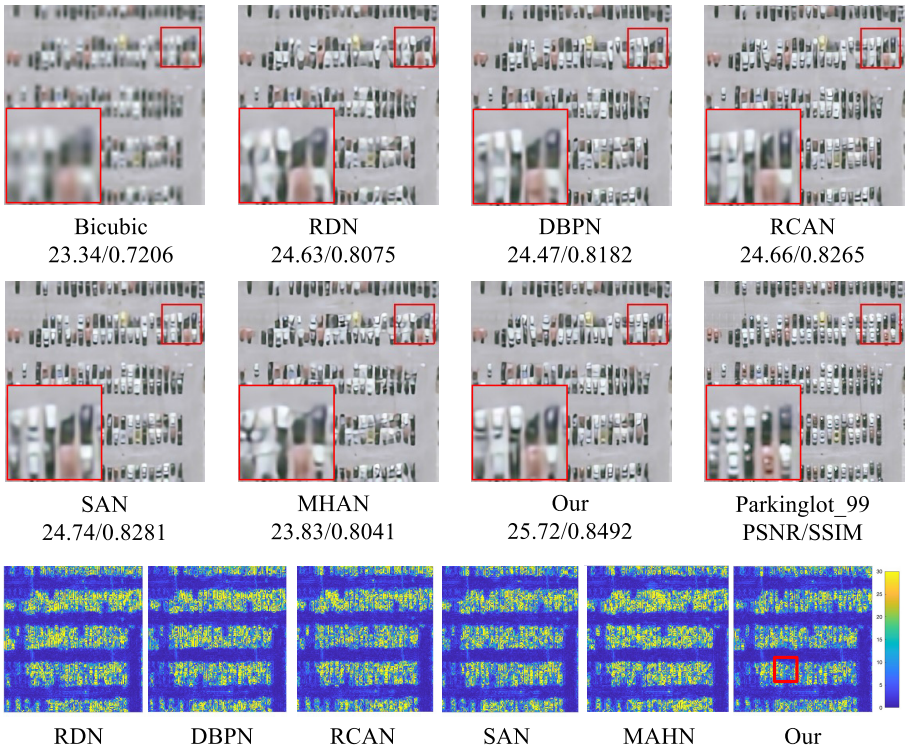


Fig. 7. Qualitative comparison of the proposed method with six counterparts on a typical satellite image pair from the UC Merced dataset with an upsampling factor of ×4. Images in the last row visualize the MSE between the SR results and the ground truth. (Zoomed-in view to see more details.)

TABLE II

QUANTITATIVE COMPARISON OF SEVEN METHODS ON THE RSCNN7 DATASET. BEST AND SECOND-BEST SCORES ARE **HIGHLIGHTED** AND UNDERLINED, RESPECTIVELY. ↑ INDICATES THAT THE LARGER THE VALUE, THE BETTER THE PERFORMANCE, AND ↓ INDICATES THAT THE SMALLER THE VALUE, THE BETTER THE PERFORMANCE

Methods	Scales	#Params/M	FLOPs/G	PSNR ↑	SSIM↑	FSIM↑	LPIPS[50] ↓	MPS [51]↑
Bicubic	× 2	-	-	32.90	0.8899	0.9922	0.2602	0.8148
RDN [52]		1.00	2.32	35.14	0.9298	0.9978	0.1779	0.8759
DBPN [53]		5.95	34.76	35.39	0.9332	0.9979	0.1759	0.8786
RCAN [48]		15.44	35.36	35.54	0.9353	0.9979	0.1632	0.8861
SAN [54]		15.67	36.10	<u>35.55</u>	<u>0.9355</u>	<u>0.9979</u>	<u>0.1630</u>	<u>0.8862</u>
MHAN [26]		11.20	26.10	35.35	0.9329	0.9979	0.1725	0.8802
Our		2.75	7.93	**35.59**	**0.9364**	**0.9981**	**0.1628**	**0.8868**
Bicubic	× 4	-	-	28.44	0.7148	0.9049	0.5747	0.5700
RDN [52]		1.15	3.73	29.86	0.7742	0.9476	0.4316	0.6713
DBPN [53]		10.43	92.68	29.83	0.7763	0.9487	0.4181	0.6791
RCAN [48]		15.59	36.77	29.75	0.7731	0.9481	<u>0.4137</u>	<u>0.6797</u>
SAN [54]		15.82	37.51	29.72	0.7727	0.9479	0.4169	0.6779
MHAN [26]		11.35	28.53	<u>29.83</u>	<u>0.7765</u>	<u>0.9488</u>	0.4226	0.6750
Our		2.75	31.74	**30.15**	**0.7835**	**0.9526**	**0.4115**	**0.6860**

Parameters (Params) and floating-point operations per second (FLOP) are tested on an LR image with 48 × 48 pixels.

100 images from each category for training and testing, respectively.

Table II reports the average performance of the proposed method and other competing deep-learning-based algorithms on the RSCNN7 dataset with scale factors ×2 and ×4, where the bold values denote the best results and underlined values represent the second-best results. The results suggest that the proposed method achieves better performance than the most competitive general/satellite image SR methods (SAN [54] and MHAN [26]) in terms of all objective evaluation indexes.

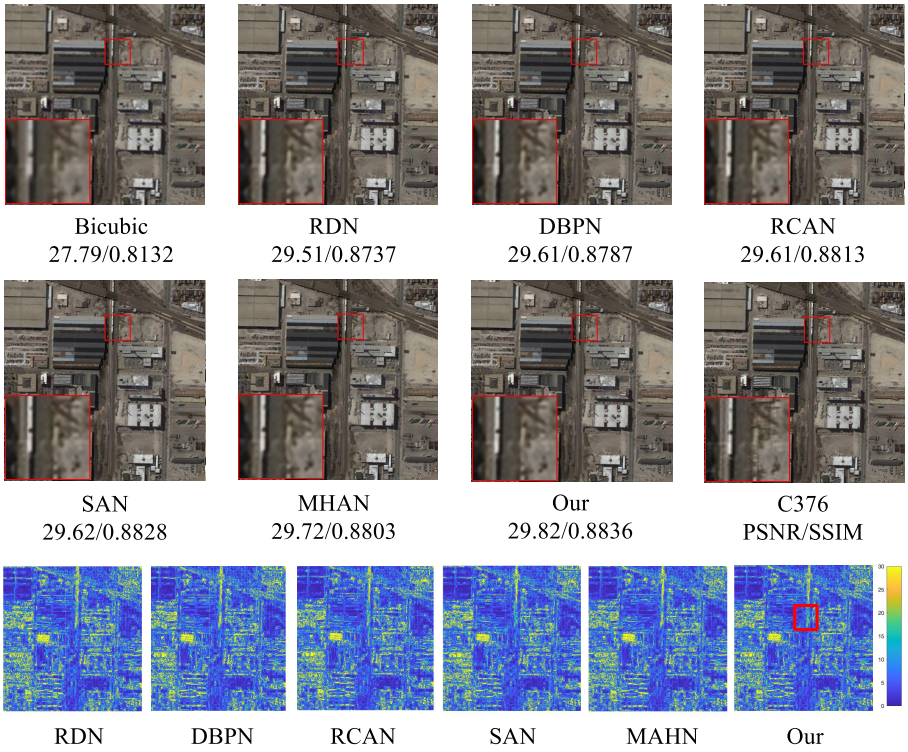


Fig. 8. Qualitative comparison of the proposed method with six counterparts on a typical satellite image pair from the remote sensing scene classification network (RSSCN7) dataset with an upsampling factor of $\times 2$. Images in the last row visualize the MSE between the SR results and the ground truth.

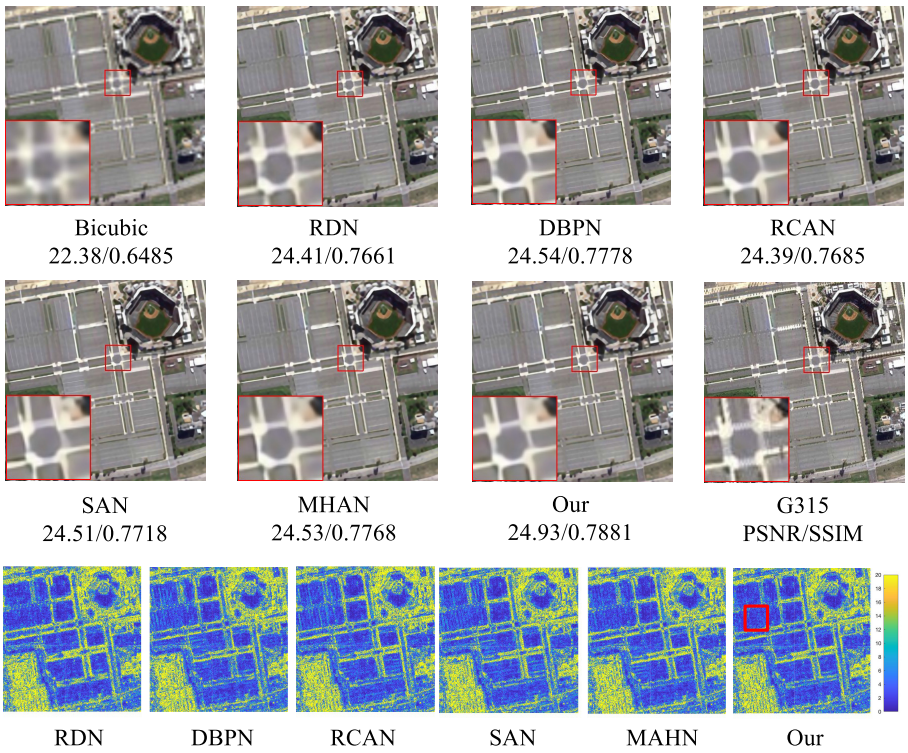


Fig. 9. Qualitative comparison of the proposed method with six counterparts on a typical satellite image pair from the RSSCN7 dataset with an upsampling factor of $\times 4$. Images in the last row visualize the MSE between the SR results and the ground truth.

The average PSNR values of the proposed method are 0.04/0.32 dB higher than that of the second-best method, with upsampling factors $t = 2$ and $t = 4$, respectively.

Figs. 8 and 9 show the reconstructed images on the RSCNN7 dataset from different competing algorithms with upsampling factors $\times 2$ and $\times 4$, respectively. Images in the

TABLE III

QUANTITATIVE COMPARISON OF SEVEN METHODS ON THE UCAS-AOD DATASET. BEST AND SECOND-BEST SCORES ARE **HIGHLIGHTED** AND UNDERLINED, RESPECTIVELY. ↑ INDICATES THAT THE LARGER THE VALUE, THE BETTER THE PERFORMANCE, AND ↓ INDICATES THAT THE SMALLER THE VALUE, THE BETTER THE PERFORMANCE

Methods	Scales	#Params/M	FLOPs/G	PSNR ↑	SSIM↑	FSIM↑	LPIPS[50] ↓	MPS [51]↑
Bicubic	× 2	-	-	34.89	0.9174	0.9408	0.1443	0.8866
RDN [52]		1.15	3.73	38.17	0.9537	0.9726	0.0723	0.9407
DBPN [53]		5.95	34.76	38.30	0.9548	0.9735	0.0727	0.9410
RCAN [48]		15.44	35.36	<u>38.39</u>	<u>0.9556</u>	<u>0.9740</u>	<u>0.0714</u>	<u>0.9421</u>
SAN [54]		15.67	36.10	38.32	0.9550	0.9737	0.0719	0.9416
MHAN [26]		11.20	26.10	38.29	0.9548	0.9735	0.0723	0.9413
Our		2.75	7.93	**38.51**	**0.9562**	**0.9742**	**0.0688**	**0.9437**
Bicubic	× 4	-	-	29.39	0.7685	0.8274	0.4403	0.6641
RDN [52]		1.15	3.73	31.44	0.8299	0.8805	0.2639	0.7830
DBPN [53]		10.43	92.68	31.69	<u>0.8362</u>	0.8851	0.2579	0.7892
RCAN [48]		15.59	36.77	<u>31.70</u>	0.8359	<u>0.8855</u>	<u>0.2532</u>	<u>0.7914</u>
SAN [54]		15.82	37.51	31.69	0.8360	0.8849	0.2549	0.7906
MHAN [26]		11.35	28.53	31.36	0.8273	0.8791	0.2548	0.7862
Our		2.75	31.74	**31.81**	**0.8385**	**0.8886**	**0.2530**	**0.7928**

Parameters (Params) and floating-point operations per second (FLOP) are tested on an LR image with 48 × 48 pixels.

last row present the MSE between the reconstructed images and the ground-truthing images. Note that the reconstructed results of RCAN [48] and SAN [54] contain notable image artifacts. The proposed method is able to capture objects' main structural information, thus resulting in well-predicted edges. Due to the fact that the RSCNN7 dataset is more challenging than the UC Merced dataset, the performance improvement of the proposed method on the RSCNN7 dataset is not as significant as the one on the UC Merced dataset.

D. Comparing With State-of-the-Arts on UCAS-AOD Dataset

The UCAS-AOD dataset is a popular dataset for remote sensing object detection. It contains 1000 images (1,280 × 659 pixels) collected from Google Earth, which mainly focuses on planes. We select 900 images and 100 images for training and testing, respectively.

Table III reports the average performance of the proposed method and other competing deep-learning-based algorithms on the UCAS-AOD dataset with scale factors ×2 and ×4, where the bold values denote the best results and underlined values represent the second-best results. It is clear that the proposed method obtains the best objective results for all evaluation indicators. Specifically, the average PSNR values of the proposed method are 0.12/0.11 dB higher than that of the second-best method, with upsampling factors $t = 2$ and $t = 4$, respectively. RCAN [48] and SAN [54] can obtain satisfactory results. However, their LPIPS/MPS values are lower than the proposed method, especially under the scenario of upsampling factor ×2.

Figs. 10 and 11 show the reconstructed images on the RSCNN7 dataset from different competing algorithms with upsampling factors ×2 and ×4, respectively. Images in the last row present the MSE between the reconstructed images and the ground-truthing images. From the reconstructed results of these methods, we notice that the proposed method is able to obtain accurate contour and reconstruct texture information of objects, which is also reflected in the error maps.

TABLE IV

RESULTS OF ABLATION EXPERIMENTS OVER RSSCN7 DATASETS WITH THE SCALE FACTOR ×4

Net_{art}	Net_{sr}	$\mathcal{L}_{ffl}$	PSNR↑	SSIM↑
✓	-	-	29.08	0.7497
-	✓	-	29.14	0.7556
✓	✓	-	29.92	0.7786
✓	✓	✓	**30.15**	**0.7835**

E. Ablation Studies

In this section, we perform the ablation studies to verify the effectiveness of the proposed artifact removal network (Net_{art}), the high-frequency generation network (Net_{sr}), and the FFL on the RSSCN7 dataset. We tabulate the objective results of the variants of the proposed framework in Table IV. We further discuss the effectiveness of the proposed components with the results.

1) Validation on Net_{art}: With the removal of the Net_{sr}, the performance reduces by 0.78 dB (Table IV). As analyzed in Section I, the edge component of the bicubic version of the LR image tends to focus on structural and artifact features, presumably resulting from the process of image degradation and upsampling. Hence, the direct prediction of the high-frequency information is artifact-prone, thus reducing the reconstruction performance.

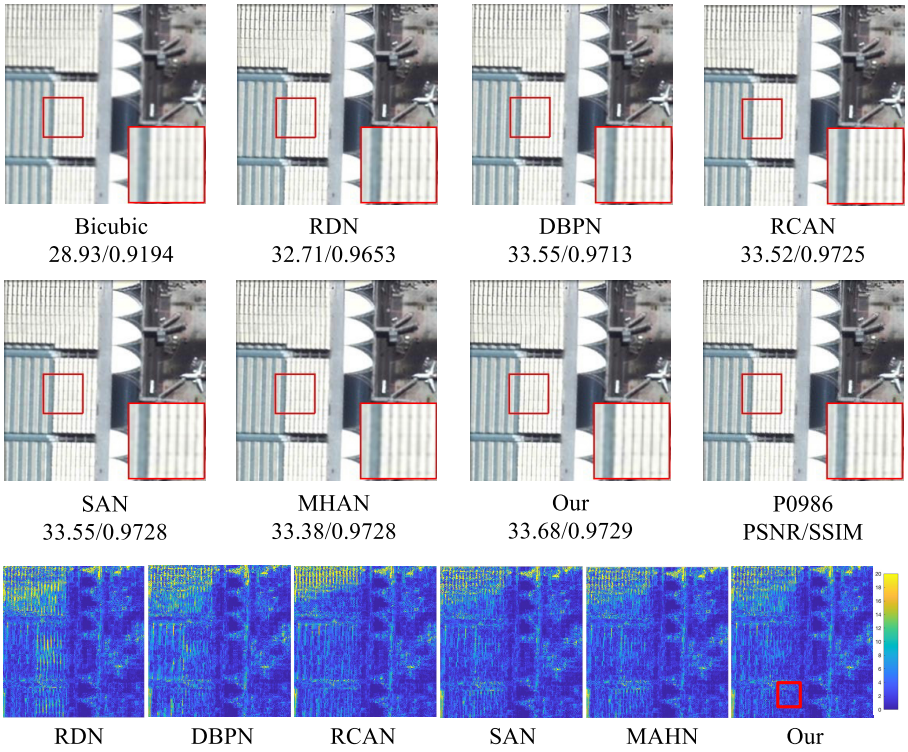


Fig. 10. Qualitative comparison of the proposed method with six counterparts on a typical satellite image pair from the UCAS-AOD dataset with an upsampling factor of $\times 2$. Images in the last row visualize the MSE between the SR results and the ground truth.

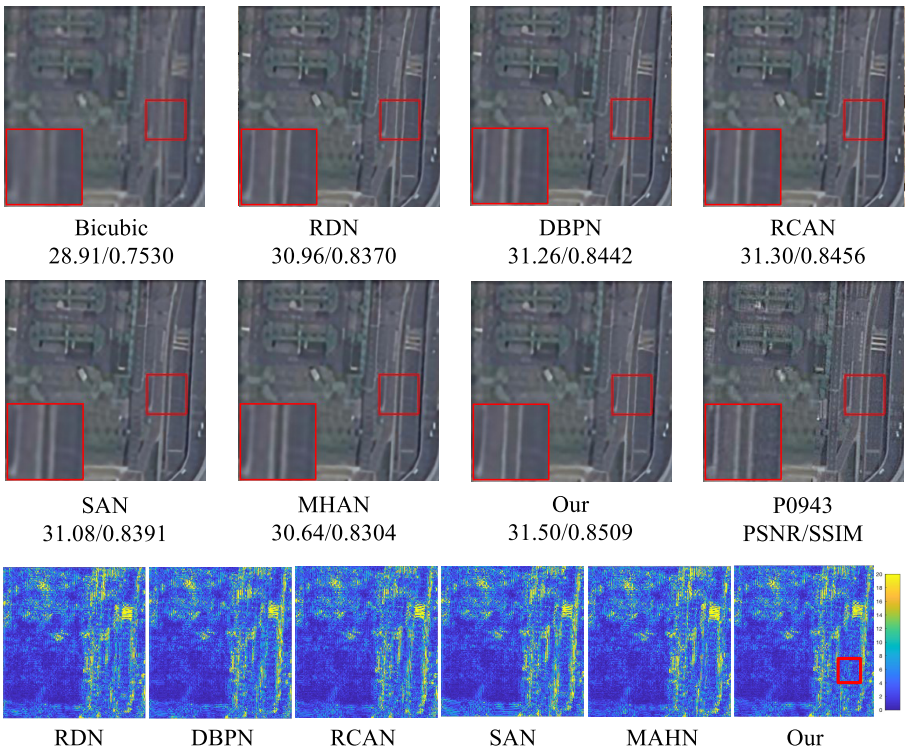


Fig. 11. Qualitative comparison of the proposed method with six counterparts on a typical satellite image pair from the UCAS-AOD dataset with an upsampling factor of $\times 4$. Images in the last row visualize the MSE between the SR results and the ground truth.

2) Validation on Net_{sr}: In this experiment, we keep the Net_{art} while remove Net_{sr} to verify the proposed artifact removal network. We note that model performance reduces by more than 0.8 dB when Net_{sr} is removed. As shown in the formula (27), the optimization of the artifact removal network is guided by the artifact information of the satellite images,

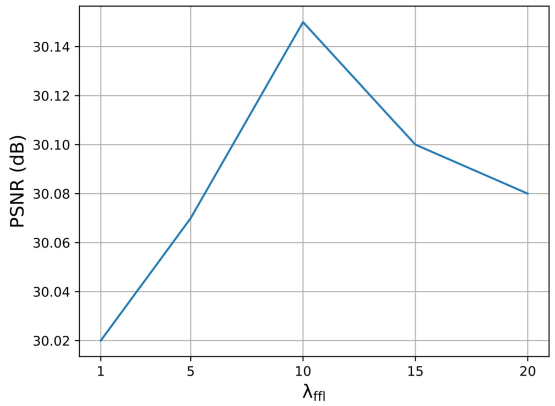


Fig. 12. Impact of λ_{ffl} in the proposed method.

leading to the smoothness of reconstructed images. Therefore, the reconstruction image of Net_{sr} is smooth, as shown in Fig. 5. This proves that the introduction of Net_{sr} benefits the reconstruction of texture and edges.

3) Validation on $\mathcal{L}_{ffl}$: To exploit the effectiveness of frequency prior, we remove the FFL loss function. We observe that the performance of the proposed method reduces when FFL loss is discarded from the model, evidenced by a reduced PSNR of 0.23 dB. FFL is expected to facilitate the alignment of textures and edges of reconstructed images and the ground-truthing images in the frequency domain.

4) Validation on λ_{ffl}: It is widely acknowledged that the hyperparameter λ_{ffl} in the loss function plays an important role in balancing the impact of the pixel loss and the frequency loss. To further validate the effectiveness of the FFL loss function, we conduct the experiments to test different settings of parameter λ_{ffl}. Fig. 12 shows the average PSNR value with different λ_{ffl} ranging from 1 to 20 on the RSCNN7 dataset with the scale factor $\times 4$. When $\lambda_{\text{ffl}} = 10$, the proposed method achieves the best performance. Excessively large values of λ_{ffl} may weaken the influence of the pixel loss, leading to more attention paid to global information, while ignoring texture information.

In order to investigate the performance of the proposed SDC, we conduct additional experiments as shown in Table V. As previously discussed, θ is the control hyperparameter that affects the performance of the algorithm. First, we replace the SDC with the convolution layer in the proposed framework to validate its validity ($\theta = 0$). Then, we train the proposed method with θ that increases from between 0 to 1. Without the SDC, the performance of the proposed method is significantly degraded (0.2 dB). With the increase of the hyperparameter θ, the proposed SDC contributes more to extracting gradient-level detailed information. Note that the proposed method achieves the best performance when $\theta = 0.8$. However, the performance decreases with the continuous increase of θ, presumably due to the fact that the gradient and content information are indispensable for SR tasks.

In addition to the above proposed strategy, we also report the results of different training steps in Table VI. "–" denotes the residual subtraction learning, and "+" denotes the residual addition learning. Pattern "1" denotes a two-stage artifact removal network while Pattern "3" denotes the utilization of a residual network. Pattern "4" can be seen as the reverse process of Pattern "2," the proposed strategy. Notably, the proposed strategy achieves the best performance, and the performance of Pattern "4" achieves the second performance.

TABLE V

RESULTS OF THE PROPOSED SDC OVER RSSCN7 DATASETS WITH THE SCALE FACTOR $\times 4$

SDC	PSNR↑	SSIM↑
$\theta = 0$	29.95	0.7803
$\theta = 0.2$	30.02	0.7818
$\theta = 0.4$	30.03	0.7817
$\theta = 0.6$	30.09	0.7824
$\theta = 0.8$	**30.15**	**0.7835**
$\theta = 1$	30.07	0.7822

TABLE VI

RESULTS ABOUT DIFFERENT TRAINING STEPS OVER RSSCN7 DATASETS WITH THE SCALE FACTOR $\times 4$

	Net_{art}	Net_{sr}	PSNR↑	SSIM↑
1	–	–	30.00	0.7810
2	–	+	**30.15**	**0.7835**
3	+	+	30.04	0.7817
4	+	–	30.07	0.7818

"–" denotes the residual subtraction learning and "+" denotes the residual addition learning.

TABLE VII

COMPARATIVE SR RESULTS ON THE MIRFLICKR-25K WITH AN UPSAMPLING FACTOR OF $\times$ 4

Method	PSNR	SSIM	FSIM
Bicubic	25.91	0.7183	0.8140
RCAN [48]	27.58	0.7768	0.8712
SAN [54]	27.60	0.7771	0.8710
MHAN [26]	27.53	0.7764	0.8684
Our	**27.74**	**0.7801**	**0.8732**

F. Model Performance on a Natural Dataset

To further evaluate the expandability of the proposed method regarding the training data, we conduct additional experiments on Multimedia Information Retrieval from FLICKR (MIRFLICKR)-25K [58], a widely used social image database. As shown in Table VII, the proposed method presents superiority over the state-of-the-art natural (RCAN [48] and SAN [54]) and satellite (MHAN [26]) SR methods in terms of PSNR, SSIM, and FSIM with an upsampling factor $\times 4$. From the visual reconstructed results in Fig. 13, we note that the proposed method can well reconstruct structures, e.g., the contours of doors and windows.

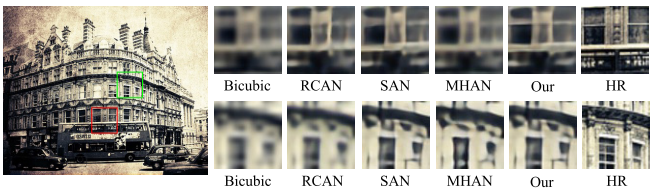


Fig. 13. SR results on the MIRFLICKR-25K dataset with an upsampling factor of $\times 4$ SR.

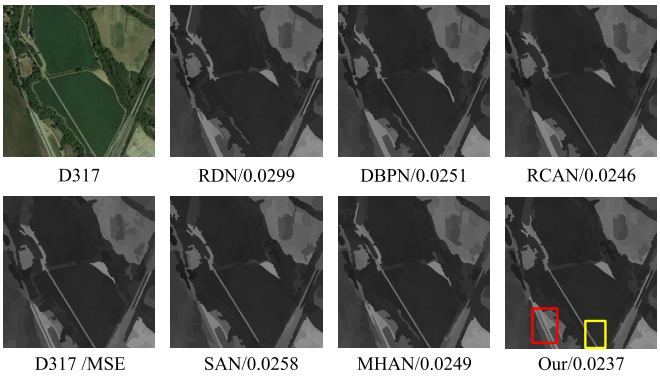


Fig. 14. Segmentation results via the spatial-spectral kernel. The colored boxes emphasize the areas where the proposed model outperforms other competing methods.

G. Performance in Segmentation Task

To further verify the impact of reconstructed images on the subsequent image segmentation task, we employ an unsupervised spatial-spectral kernel (with default parameters) [59] as the segmentation filtering for satellite image semantic segmentation. As shown in Fig. 14, the colored boxes highlight the areas where the proposed model outperforms other competing methods. Note that the bridge crossed the long river (see the yellow box) in the results of DBPN, MHAN, and the proposed method can be accurately delineated. For the two roads in the lower-left corner, only the proposed method can reconstruct their shapes, suggesting its superior performance compared with other state-of-the-art algorithms. In order to quantitatively evaluate the segmentation results, we employ the MSE to measure the distance between the predicted results of SR and HR images. It is clear that the proposed method achieves optimal quantitative results (i.e., 0.00237).

V. Conclusion

In this work, we develop an end-to-end deep CNN framework that decomposes a HR image into three components, i.e., LR, artifact, and high-frequency information. The proposed method includes two major subnetworks, i.e., an artifact removal network and a high-frequency generation network. We employ the SDC block to exploit the structural information hidden in the ringing effect for the artifact prediction. Considering the complementary nature between the artifact image and the residual image, we introduce an SSC block to connect these two networks. Numerous experiments and ablation studies demonstrate that the proposed method exhibits state-of-the-art performance. Future works are expected to employ LR images as input and reduce the computational demand.

References

[1] D. Li, J. Shan, Z. Shao, X. Zhou, and Y. Yao, "Geomatics for smart cities–concept, key techniques, and applications," *Geo-Spatial Inf. Sci.*, vol. 16, no. 1, pp. 13–24, 2013.

[2] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-Spatial Inf. Sci.*, vol. 25, no. 2, pp. 278–294, 2022.

[3] Q. Zhuang *et al.*, "Evolution of soil salinization under the background of landscape patterns in the irrigated northern slopes of Tianshan mountains, Xinjiang, China," *Catena*, vol. 206, Nov. 2021, Art. no. 105561.

[4] Q. Zhuang *et al.*, "Unequal weakening of urbanization and soil salinization on vegetation production capacity," *Geoderma*, vol. 411, Apr. 2022, Art. no. 115712.

[5] C. Dang *et al.*, "Assessment of the importance of increasing temperature and decreasing soil moisture on global ecosystem productivity using solar-induced chlorophyll fluorescence," *Global Change Biol.*, vol. 28, no. 6, pp. 2066–2080, Mar. 2022, doi: 10.1111/gcb.16043.

[6] H. Yu, J. Wang, Y. Bai, W. Yang, and G.-S. Xia, "Analysis of large-scale UAV images using a multi-scale hierarchical representation," *Geo-Spatial Inf. Sci.*, vol. 21, no. 3, pp. 33–44, 2018.

[7] S. Zhang, Z. Shao, X. Huang, L. Bai, and J. Wang, "An internal-external optimized convolutional neural network for arbitrary orientated object detection from optical remote sensing images," *Geo-Spatial Inf. Sci.*, vol. 24, no. 4, pp. 654–665, Oct. 2021.

[8] Z. Wang, J. Chen, and S. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Mar. 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] W. Yang *et al.*, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5895–5907, Dec. 2017.

[11] Q. Mao, S. Wang, S. Wang, X. Zhang, and S. Ma, "Enhanced image decoding via edge-preserving generative adversarial networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[12] 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, Jun. 2019.

[13] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014.

[14] 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. 2015.

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

[16] 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.

[17] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[18] X. Qin, X. Gao, and K. Yue, "Remote sensing image super-resolution using multi-scale convolutional neural network," in *Proc. 11th U.K.-Europe-China Workshop Millim. Waves Terahertz Technol. (UCMMT)*, Sep. 2018, pp. 1–3.

[19] Z. Tu, W. Xie, J. Dauwels, B. Li, and J. Yuan, "Semantic cues enhanced multimodality multistream CNN for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1423–1437, May 2019.

[20] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial–spectral prior for super-resolution of hyperspectral imagery," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1082–1096, 2020.

[21] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, and J. Ma, "Enhanced image prior for unsupervised remoting sensing super-resolution," *Neural Netw.*, vol. 143, pp. 400–412, Nov. 2021.

[22] T. Lu, J. Wang, Y. Zhang, Z. Wang, and J. Jiang, "Satellite image super-resolution via multi-scale residual deep neural network," *Remote Sens.*, vol. 11, no. 13, p. 1588, Jul. 2019.

[23] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3512–3527, Jun. 2019.

[24] X. Dong, X. Sun, X. Jia, Z. Xi, L. Gao, and B. Zhang, "Remote sensing image super-resolution using novel dense-sampling networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1618–1633, Feb. 2021.

[25] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote sensing image superresolution using deep residual channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019.

[26] D. Zhang, J. Shao, X. Li, and H. T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5183–5196, Jun. 2021.

[27] 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.

[28] P. Lei and C. Liu, "Inception residual attention network for remote sensing image super-resolution," *Int. J. Remote Sens.*, vol. 41, no. 24, pp. 9565–9587, Dec. 2020.

[29] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[30] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3633–3643, May 2020.

[31] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.

[32] 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.

[33] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, and X. Guo, "SDPNet: A deep network for pan-sharpening with enhanced information representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4120–4134, May 2021.

[34] J. Wang, Z. Shao, X. Huang, T. Lu, and R. Zhang, "A dual-path fusion network for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[35] Y. Qu, H. Qi, C. Kwan, N. Yokoya, and J. Chanussot, "Unsupervised and unregistered hyperspectral image super-resolution with mutual Dirichlet-net," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[36] C. Grossmann, H.-G. Roos, and M. Stynes, *Numerical Treatment of Partial Differential Equations*, vol. 154. Springer, 2007.

[37] C. Jing, D. Zhang, P. Ni, K. Zhang, and L. Yang, "Difference networks and second-order difference networks," in *Proc. 4th Int. Conf. Syst. Informat. (ICSAI)*, Nov. 2017, pp. 1373–1377.

[38] M. Sarıgül, B. M. Ozyildirim, and M. Avci, "Differential convolutional neural network," *Neural Netw.*, vol. 116, pp. 279–287, Aug. 2019.

[39] Z. Yu *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5295–5305.

[40] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, "Dual-cross central difference network for face anti-spoofing," 2021, *arXiv:2105.01290*.

[41] L. Liu, L. Zhao, Y. Long, G. Kuang, and P. Fieguth, "Extended local binary patterns for texture classification," *Image Vis. Comput.*, vol. 30, no. 2, pp. 86–99, Feb. 2012.

[42] Z. Su *et al.*, "Pixel difference networks for efficient edge detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5117–5127.

[43] J. Klimack, "A study on different architectures on a 3D garment reconstruction network," M.S. thesis, Dept. Math. Inform., Universitat Politècnica de Catalunya, Barcelona, Spain, 2021.

[44] S. Miao, Y. Hou, Z. Gao, M. Xu, and W. Li, "A central difference graph convolutional operator for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4893–4899, Jul. 2022.

[45] S. Zabihi, H. Rezazadegan Tavakoli, and A. Borji, "A compact deep architecture for real-time saliency prediction," 2020, *arXiv:2008.13227*.

[46] Y. Zhao, B. Zou, F. Yang, L. Lu, A. N. Belkacem, and C. Chen, "Video-based physiological measurement using 3D central difference convolution attention network," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–6.

[47] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, "MTD-net: Learning to detect deepfakes images by multi-scale texture difference," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4234–4245, 2021.

[48] 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.*, 2018, pp. 286–301.

[49] 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.

[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[51] M. El Helou *et al.*, "AIM 2020: Scene relighting and illumination estimation challenge," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 499–518.

[52] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[53] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[54] 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. 11065–11074.

[55] D. Fuoli, L. Van Gool, and R. Timofte, "Fourier space losses for efficient perceptual image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2360–2369.

[56] 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. (GIS)*, 2010, pp. 270–279.

[57] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[58] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr. (MIR)*, 2008, pp. 39–43.

[59] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial–spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognit.*, vol. 45, no. 1, pp. 381–392, 2012.

Jiaming Wang received the B.S. and master's degrees from the College of Post and Telecommunication, Wuhan Institute of Technology, Wuhan, China, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree under the supervision of Prof. Zhenfeng Shao with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan.

His research field includes image/video processing and computer vision.

Zhenfeng Shao received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2004.

Since 2009, he has been a Full Professor with LIESMARS, Wuhan University. He has authored or coauthored over 50 peer-reviewed articles in international journals. His research interests include high-resolution image processing, pattern recognition, and urban remote sensing applications.

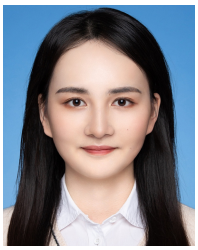
Dr. Shao was a recipient of the Talbert Abrams Award for the Best Paper in Image matching from the American Society for Photogrammetry and Remote Sensing in 2014 and the New Century Excellent Talents in University from the Ministry of Education of China in 2012. He has served as an Associate Editor of the Photogrammetric Engineering and Remote Sensing (PE and RS) specializing in smart cities, photogrammetry and change detection since 2019.

Xiao Huang received the B.S. degree from Wuhan University, Wuhan, China, in 2015, the master's degree from the Geographic Information Science and Technology, Georgia Institution of Technology, Atlanta, GA, USA, in 2016, and the Ph.D. degree in geography from the University of South Carolina, Columbia, SC, USA, in 2020.

He is currently an Assistant Professor with the Department of Geosciences, University of Arkansas, Fayetteville, AR, USA. His research interests cover GeoAI, deep learning, and human-environmental interactions.

Tao Lu (Member, IEEE) received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China, in 2003 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, in 2013.

He was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, from 2015 to 2017. He is an Associate Professor with the School of Computer Science and Engineering, Wuhan Institute of Technology and a Research Member with the Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan. His research interests include image/video processing, computer vision, and artificial intelligence.



Ruiqian Zhang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2021.

She is currently a Post-Doctoral Fellow with the Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping, Beijing, China. Her research interests include image processing, pattern recognition, and remote sensing.

Yong Li is currently pursuing the M.A. degree under the supervision of Prof. Zhenfeng Shao with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

His research interests are urban remote sensing and computer vision.