Remote Sensing Image Super-Resolution With Residual Split Attention Mechanism

Xitong Chen, Yuntao Wu, Tao Lu, *Member, IEEE*, Quan Kong, Jiaming Wang, and Yu Wang

***Abstract*—Recently, deep-learning-based methods have become the current mainstream of remote sensing image super-resolution (SR) due to their powerful fitting ability. However, they are still unsatisfactory in large-scale factor SR scenarios. The more complicated information distribution of images further increases the difficulty of reconstruction. In this article, we propose a novel residual split attention group (RSAG) to maintain the overall structural and the local details simultaneously. Specifically, an upscale module that makes the network jointly consider hierarchical priors, which assists in the prediction of high-frequency information, and a residual split attention module to adaptively explore and exploit the global structure information in low-level feature space. In addition, an artifact removal strategy is proposed to reduce excessive artifacts and further boost the performance. By progressively connecting the above modules and incrementally fusing the multilevel intermediate feature maps, the fidelity of high-frequency detail information is improved. Finally, we propose a residual split attention network by stacking several RSAGs for reconstructing high-resolution remote sensing images. Extensive experiment results demonstrate that the proposed approach achieves better quantitative metrics and visual quality than the state-of-the-art approaches.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), remote sensing image, super-resolution (SR).**

I. Introduction

AS AN important means of earth observation, the remote sensing image is widely used in a variety of fields, including mineral resources, environmental monitoring, public safety, and military applications. The spatial resolution is the most significant indicator of the quality of the satellite image. However, the spaceborne imaging systems are often affected by the complex imaging environment, resulting in low spatial resolution of the acquired images. Therefore, the captured satellite images may not be accurate enough for advanced remote sensing applications such as object detection [1], image segmentation [2], etc. Image super-resolution (SR) is an algorithmic technique for producing a potentially high-resolution (HR) image from a given low-resolution (LR) image.

With the rapid development of satellite photogrammetry, it is very urgent to develop efficient and high-precision satellite image SR methods. Tsai [3] pioneered the use of fusing complementary information for satellite image SR tasks and utilized the complementary information between different frameworks to reconstruct HR remote sensing images. Recently, SR algorithms have been effectively utilized to improve image resolution and quality, which are widely used in preprocessing techniques for remote sensing image analysis [4]. The current SR methods can be categorized as interpolation-, reconstruction-, and learning-based methods. The interpolation-based methods [5], [6] are a kind of noniterative framework, whose core idea is to align the LR with the HR remote sensing image and apply nonuniform interpolation to obtain the value of each pixel corresponding to the HR remote sensing image grid. Reconstruction-based methods [7], [8] typically entail converting HR images to LR images by using downsampling, establishing correspondence by studying the performance of HR detail information under LR conditions, and ultimately expressing this relationship through modeling. One of the classical remote sensing SR reconstruction algorithms is the hidden Markov chain model proposed by Li et al. [9]. Because this model relies on accurate subpixel accuracy estimates, reconstructed remote sensing images may be severely lacking in high-frequency detail information and can only boost small magnifications [10].

Influenced by the speed development of machine learning, the deep learning-based satellite image SR approaches, gradually become a mainstream research direction. The deep learning-based algorithms show strong feature representation ability, which can be used to learn the nonlinear function by convolutional neural networks (CNNs) and achieve satisfactory results. As a result, more and more CNN-based remote sensing SR methods are being proposed by scholars. But most studies [11], [12], [13], [14], [15] on the SR of remote sensing images have focused on small magnification factors, and increased the resolution by adding an upsampling layer. Few studies have attempted to solve the reconstruction problem with a large magnification factor. Such as, Pan et al. [16] utilized the backprojection strategy to handle of the dependency between LR and HR more completely. Dong et al. [17] proposed a dense-sampling framework that

Manuscript received 1 March 2023; revised 12 May 2023; accepted 8 June 2023. Date of publication 21 June 2023; date of current version 12 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61771353 and Grant 62072350, in part by the Three Gorges Laboratory Open Fund of Hubei Province under Grant SC215001, in part by the Key R&D Program of Hubei under Grant 2022BAA052, and in part by the Research Plan Project of Hubei Provincial Department of Education under Grant B2022062. *(Corresponding author: Yuntao Wu.)*

Xitong Chen, Tao Lu, Quan Kong, and Jiaming Wang are with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China (e-mail: chenxitong21@gmail.com; lutxyl@gmail.com; witkongquan@wit.edu.cn; wjmecho@whu.edu.cn).

Yuntao Wu is with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China, and also with the HuBei Three Gorges Laboratory, Yichang 443007, China (e-mail: ytwu@sina.com).

Yu Wang is with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: wy2022@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3287894

reused an upscaler to upsample low-dimensional features via a dense-sampling mechanism for investigating large-scale factor SR reconstruction. Although the existing methods, RDBPN [16] and DSSR [17], have attempted to densely connect the feature by upsampling modules to address the reconstruction challenge at high magnification levels (e.g., $\times 8$), which used the same weight size to learn remote sensing feature maps of different regions and lacked discriminative ability across feature channels. During the application process, the connection between the hierarchical features is lost, resulting in the loss of intermediate information. In addition, due to the limited a priori factors available in LR space under severe upscaling circumstances, reliable prediction of local features becomes extremely challenging. The additional prior information means more computational overhead and increases the training difficulty, thus negatively affecting the subsequent reconstruction process.

To meet the aforementioned challenges, we propose an efficient residual split attention network (RSAN). First, we propose a multipath residual split attention (RSA) mechanism to promote the internal correlation of features by splitting and fusing different channel dimensions, which ensures the method pays more attention to detail-rich regions and focuses less on parts that are not well-informed. Then, we design an upscale module to learn the hierarchical prior information in an HR potential subspace to help the prediction of high-frequency information and a residual split attention module (RSAM) with the proposed RSA and a downscale operation to explore and exploit global information in LR potential subspace. In addition, an artifact removal strategy is proposed in the upscale module to reduce excessive artifacts for better large-scale factor SR. The upscale module and RSAM are combined to form the residual split attention group (RSAG), which is introduced to simultaneously enhance the consistency of global structure and the fidelity of local detail restoration by fusing multilevel and multipath features. In each RSAG module, we adopt a dense connection to conduct the residual features fusion (RFF), which across different levels, reducing feature redundancy and promoting information exchange within and between modules. Finally, we propose a RSAN to cascade several RSAGs for reconstructing HR remote sensing images. The main contributions are as follows.

1) We propose a residual split attention network for the single remote sensing image SR, which can better balance the model size and achieve superior results on two publicly available datasets even under the large-scale factor.
2) We present the RSAM to assist the network in focusing training on detail-rich regions while paying less attention to parts that are not well-informed by splitting and fusing the intermediate residual feature maps from different channel dimensions and strengthening the representation capacity of the network.
3) To ensure both the global structural consistency and the local detail restoration fidelity are fully maintained, the RSAG is proposed to use an upscale module to jointly consider the hierarchical prior information and connect with an RSAM for adaptive weighted fusion of multipath information, which enables the network to be more accurate for reconstruction by exploiting different dimensional information.

The rest of this article is organized as follows. Section II describes the related work. Section III illustrates detailed description of the proposed method. Section IV provides experimental results. Finally, Section V concludes this article.

II. Related Work

A. CNN-Based Image Super-Resolution

Recently, CNN-based methods demonstrate excellent performance in various computer vision tasks [18], [19], [20] because of their robust feature representation capabilities. In 2015, Dong et al. [21] first proposed three-layers CNN framework. Kim et al. [22] introduced the concept of residual networks [23] can effectively build deeper networks and converge faster. Lim et al. [24] improved a enhanced residual SR network (EDSR) based on ResNet [25] blocks, which saved space by eliminating unnecessary modules from the traditional residual network and further expanded the size of the model to enhance the network expression ability. Lai et al. [26] introduced the Laplacian pyramid framework, which can predict residuals from coarse to fine. Haris et al. [27] introduced the deep back-projection network (DBPN) that can fully exploit the interdependence of HR and LR pairings by cascading several up- and down-sampling blocks to better learn high-resolution features and achieve good performance, particularly on large-scale factor. Considering the correlation between the channels, Zhang et al. [28] presented a very deep residual channel attention network (RCAN), which can have targeted extraction of the high-frequency component, by rescaling the channelwise features to focus on the image edge texture. To improve feature expression ability, Dai et al. [29] proposed a second-order attention network (SAN) to generate discriminative features and information. Lu et al. [30] designed a multiscale information polymerization network, which addressed the problem of limited representation ability of reconstructed networks caused by the lack of consideration of the potential relationship between multiscale features in existing CNN-based SISR methods. By studying image sparsity to accelerate the inference efficiency of the network, Wang et al. [31] designed a sparse mask framework to identify different regions by using spatial and channel mask learning to mark unimportant regions that can reduce redundant computations while maintaining good performance.

The SR method described above is aimed at general images. Due to the wide range of satellite images, the spatial distribution of remote sensing images is complicated. Thus, the targets to be recovered often cover only a few pixels in the image, and the pixel differences between different types of targets are small. Therefore, deep learning-based methods designed for general images cannot effectively process satellite images due to their inability to retrieve the potential high-frequency information contained in satellite images, especially with large-scale sampling.

B. Satellite Images Super-Resolution

In remote sensing image applications, recovering HR images with clear texture details is indispensable for many tasks, because satisfactory application results cannot be obtained with

only a small amount of feature information provided by LR images. Liebel and Körner [11] were the first to apply the SRCNN [21] to satellite images SR. Considering the satellite image SR method cannot directly train by the natural images, so the authors produced a remote sensing dataset using SENTINEL-2 images to relearn the mapping relationship. Lei et al. [32] designed a multifork structured framework to learn the multiscale representation ability, which combined shallow and deep feature mappings to complete the interaction of network information to better guide the reconstruction. Qin et al. [33] introduced a multiscale network based on GoogLeNet [34] that extracted image features with multiscale kernels and obtained more comprehensive depth features after concatenating each channel feature to improve the SR effect.

Inspired by the successful application of knowledge distillation [35], [36], [37] in computer vision tasks, Jiang et al. [38] constructed a distillation framework to distill and compensate feature maps at various stages for high-frequency information enhancement. Ma et al. [39] devised a approach to simplify the training stage by the wavelet transform, which combines global with local residual learning to alleviate the problem of gradient disappearance. Gu et al. [40] developed a deep residual attention strategy, which used a residual attention block to adjust the weight of feature maps and improve the representation ability. Huan et al. [41] proposed a pyramidal multiscale residual framework to enhance the power that detect contextual information. Lu et al. [42] proposed a novel structure-texture parallel embedding (SPE) method, which utilized both global structural information and local texture information in the upscaling process to guide the reconstruction results. Wang et al. [43] designed a novel satellite SR framework to transform HR images into LR, artifact, high-frequency information and introduced a self-adaption difference convolution module to better recover remote sensing images.

C. Neural Attention Mechanism

The neural attention mechanism can focus on important region with limited resources and become a popular research topic. It originated from the exploration of the human visual mechanism. Human vision tends to focus on the salient areas while ignoring the information-poor parts, and a neural attention mechanism can help neural networks focus on important feature information while suppressing useless feature representations and improving information processing efficiency. Haut et al. [44] introduced the attention mechanism into the SR tasks to learn the mapping function between texture components, enhance the high-frequency information of the image, and suppress the low-frequency information. Dong et al. [45] designed a multiperception learning framework to perform multilevel information adaptive weighted fusion for reconstruction. Further, Zhang et al. [46] proposed the mixed high-order attention mechanism (MHAN), which applied weights to different levels of convolution in the feature extraction stage to retain more important information, and added frequency-aware connection in the feature refinement stage to fuse and refine the features of different depths through the high-order attention module. Li et al. [47] introduced an adaptive weighted attention network that integrates an adaptive weighted channel attention module and a patch-level second-order nonlocal module to capture interdependencies among intermediate features and enhance feature representations. To address the challenge of satellite images with large difference in scene and image size, Zhang et al. [48] proposed a multiscale attention network for features extracting that used the channel attention mechanism to fuse multiscale features and assigned models for the satellite images reconstruction. This method obtains good results, but the number of models and parameters increases significantly. Although the above attention mechanisms can enhance the network's learning of important features, they lack the ability to discriminatively learn different spatial regions of the same feature. Lei and Liu [49] utilized the inception module [34] to extract scale-invariant features and combined the channel and spatial attention mechanisms to distinguish important features, which allocated attention to different regions of each feature map and made the network perform more comprehensive discriminative learning of remote sensing features. To overcome the bottleneck of low accuracy in the existing unsupervised SR methods, Li et al. [50] proposed an unsupervised super-resolution architecture that included the masked transformer to extract latent hyperspectral characteristics for realistic restoration of hyperspectral images, with strong constraints incorporated into the framework. They also introduced a dual spectralwise multihead self-attention mechanism to address the limitations of traditional CNN-based models and enhance the robustness of the model.

III. Proposed Method

The structure of RSAN is described first in this section. Then, we elaborate the proposed residual split attention group, which is composed of the upscale and residual split attention modules, respectively. Finally, we introduce the loss function. In RSAN, we let $I_{LR} \in \Re^{h\times w\times c}$ and $I_{HR} \in \Re^{rh\times rw\times c}$ be the LR and ground truth images, respectively, where h, w, and c denotes the height, the width, and the channel number of the LR image. r represents the scale factor. In addition, let $Conv(n_k, n_f, n_c)$, $PwConv(n_k, n_f, n_c)$, $DwConv(n_k, n_f, n_c)$ and $DeConv(n_k, n_f, n_c)$ indicate the standard convolutional, pointwise convolutional, depthwise convolutional, and deconvolutional layers, where n_k, n_f, and n_c denote the filter size, the number of input channels, and the number of output channels, respectively.

A. Network Architecture

Fig. 1 shows the structure of RSAN. The proposed method consists of four components: coarse feature extraction part, residual split attention group, multilevel features fusion module, and reconstruction module. The coarse feature F_C is extracted in the RSAN initial part from the input LR remote sensing image I_{LR}, as

$$F_C = H_C(I_{LR}) \tag{1}$$

where $H_C(\cdot)$ is the coarse feature extracting operation with one inverted residual block and one $Conv(3, 64, 64)$ layer. In the

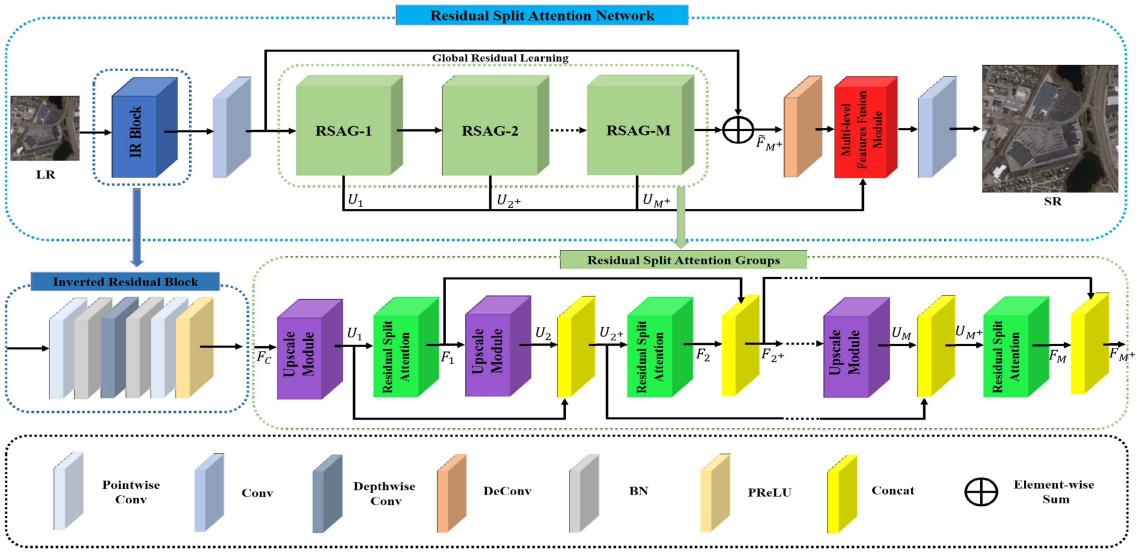

Fig. 1. Network architecture of RSAN.

middle part F_C is used for residual split attention group deep feature extraction, the mth group RSAG-m extracts the deep feature map F_{m^+} as follows:

$$F_1 = H_{RSAG,m}(F_C) \quad m = 1 \tag{2}$$

$$F_{m^+} = H_{RSAG,m}(F_{(m-1)^+}) \quad m = 2, \ldots, M \tag{3}$$

where $H_{RSAG,m}(\cdot)$ represents the mth RSAG, the specific structure will be detailed later and the number of m settings will be described in the ablation study section. Then, the global residual learning is introduced into the last RSAG output, so the input of the last deconvolutional layer is defined as follows:

$$\tilde{F}_{M^+} = H_{Sum}(F_{M^+}, F_C) \tag{4}$$

where H_{Sum} refers to elementwise sum operation. After the last RSAG, we feed $\tilde{F}_{M^+}$ into the deconvolutional [51] layer to obtain high resolution feature map $\tilde{U}_{M^+}$ and aggregate the previous multilevel HR feature maps $(U_1, U_{2^+}, \ldots, U_{M^+})$ in the multilevel feature fusion module to estimate the reconstructed HR image. The SR can be described as

$$I_{SR} = H_{Rec}\left(concat\left(U_1, U_{2^+}, \ldots, U_{M^+}, \tilde{U}_{M^+}\right)\right) \tag{5}$$

where $H_{Rec}(\cdot)$ use $Conv(3, 64 * (m+1), 3)$ as reconstruction and $concat(\cdot)$ represents the concatenation function.

B. Residual Split Attention Groups

Previous attention-based methods [45], [46], [48] only used multilevel residual blocks for refinement to generate richer hierarchical features. However, the lack of information interaction between different spaces and channels in the single-sized receptive field residual blocks. Inspired by ResNeSt [52], we propose the RSAG to excavate the internal relevance of features by multichannel splitting and classification of feature channel dimensions, and focus on detail-rich regions and pay less attention to parts that are not well-informed. Compared to ResNeSt, our proposed RSAG first learns the hierarchical features in a high-resolution potential subspace to improve the network's prediction of high-frequency information. This procedure has the additional advantage of an artifact removal operation that effectively reduces excessive artifacts, thereby achieving better large-scale factor SR. Then, we use a depthwise convolutional operation along with the residual split attention mechanism to explore and exploit global information in the LR potential subspace. Differing from ResNeSt, RSAN continuously projects the feature space across different dimensions to simulate the degradation level of remote sensing images at different stages and better learn high-resolution components. More importantly, RSAN is specifically designed to enhance the representation of hierarchical features and is an efficient remote sensing image SR network that progressively restores details from coarse to fine.

As shown in Fig. 2, the RSAG is made up of the upscale and the residual split attention module. The upscale module can upsample the coarse feature F_C, then we set the $DeConv(n_k, n_f, n_c)$ and feature extraction processes as the upscale module for RSAN. First, the upscale module maps the coarse feature F_C to an intermediate HR map $U_{1,0}$ via one deconvolutional layer $DeConv(8, 64, 64)$ with an upsampling factor of $r = 4$. When the $r = 8$, k is set to 12. Then, $U_{1,0}$ is mapped back to obtain the LR feature map L_1 through one pointwise convolutional layer $PwConv(1, 64, 64)$, one depthwise convolutional layer $DwConv(3, 64, 64)$ and one pointwise convolutional layer $PwConv(1, 64, 64)$. Subsequently, we introduce an artifact removal operation to utilize the structure prior in LR potential subspace and estimate the artifact residual feature map a. The artifact residual feature map a between the input LR F_C and the learned L is computed by a deconvolutional layer $DeConv(8, 64, 64)$ to get the HR map $U_{1,1}$. The upscale module output U_1 is computed by summing the intermediate HR map $U_{1,0}$ and $U_{1,1}$. Then, U_1 is fed into the proposed RSAM. The local residual learning feature U_1 is passed through a depthwise convolutional layer $DwConv(3, 64 * m, 64)$ to obtain $Split$.

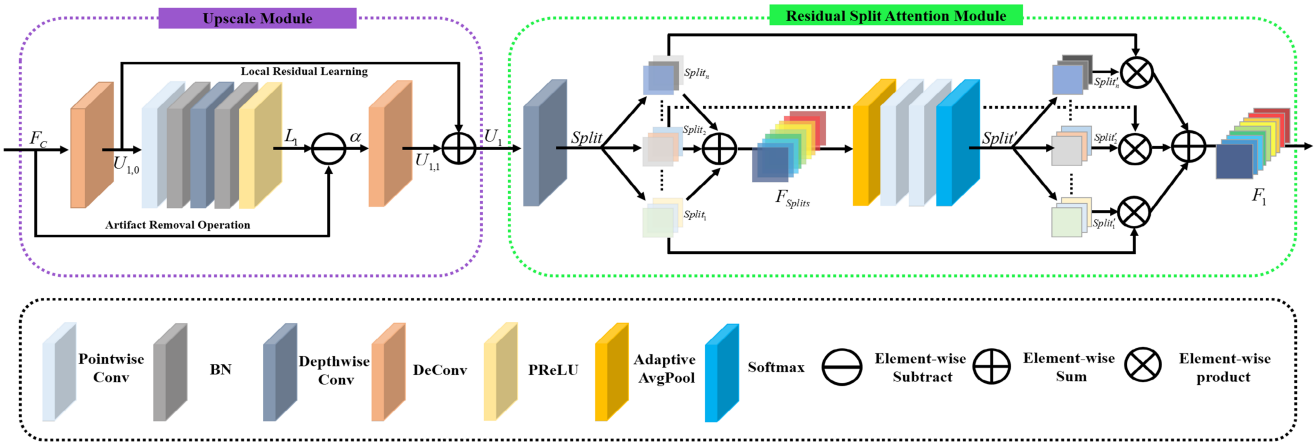

Fig. 2. Network architecture of RSAG.

Subsequently, $Split$ is split into n separate splits through each channel, where the number of output channels for each split is defined as c, resulting in a value of $\frac{c}{n}$ output channels for each split. Then, an elementwise sum operation is performed on each split

$$F_{Splits} = H_{Sum}\left(Split_1, Split_2, \ldots, Split_n\right) \tag{6}$$

where $Split_n$ represents the nth split by the last division operation. The F_{Splits} is passed through an adaptive average pooling layer and two pointwise convolutional layers, and then input to a n-softmax function based on the feature detail richness of each previous split, resulting in the latest $Split'$. Therefore, $Split'$ can be formulated as follows:

$$Split' = H_S(PwConv(PwConv(H_{Avg}(F_{Splits})))) \tag{7}$$

where H_{Avg} denotes the adaptive average pooling operation and H_S represents the softmax function. We split $Split'$ into n splits $(Split'_1, Split'_2, \ldots, Split'_n)$ again. Then, each latest $Split'_n$ is multiplied by the previous corresponding n split $Split_n$ using the product operation of the elements, respectively. By the operation of an elementwise product, the latest n splits is multiplied by the corresponding previous n splits. Thus, the internal correlation of features is improved by using multipath channel information, and the network can focus on the restoration of global structure with RSA. After the second feature split, we use the elementwise sum operation to merge each path split as the output of the RSAM, which is denotes as F_{RSA}. The 1st RSAM ouput $F_{RSA,1}$ is defined as follows:

$$\begin{aligned} F_{RSA,1} &= F_1 = H_{RSA,1}(U_1) \\ &= H_{Sum}\left(H_{Ep}\left(Split_1, Split'_1\right), \ldots,\right. \\ &\quad \left. H_{Ep}\left(Split_n, Split'_n\right)\right) \end{aligned} \tag{8}$$

where $H_{RSA,1}$ denotes the output of 1th RSAM, the H_{Ep} represents an elementwise product operation.

The upscale module and RSAM are connected with each other to constitute the RSAG, the Fig. 1 depicts the entire RSAGs structure. In the lower right part of Fig. 1, purple and green cubes represent the upscale module and RASM, respectively. The yellow cube indicates the operation of concatenating feature maps along the channel dimension. The overall construction of mth RSAG is described in detail below. The coarse feature F_C is first processed by the upscale module

$$U_1 = H_{Up,1}(F_C) \tag{9}$$

where $H_{Up,1}$ denotes the operation of the first upscale module. Then an RASM and upscale module generate initial level features F_1 and U_2 as follows:

$$F_1 = H_{RSA,1}(U_1) \tag{10}$$

$$U_2 = H_{Up,2}(F_1) \tag{11}$$

the dense connected structure [53] is used to fully utilize the different hierarchical features, in which each upscale module output aggregates feature maps from all previous upscale modules. When the group number $m \geq 2$, the concatenate module is placed after the every upscale module and RSAM, the input to the mth RSAM can be represented as follows:

$$U_{2^+} = concat(U_2, U_1) \quad m = 2 \tag{12}$$

$$U_{m^+} = concat(H_{Up,m}(F_{(m-1)^+}), U_{(m-1)^+}) \quad m = 3, \ldots, M \tag{13}$$

the input to the mth upscale module can be denotes as follows:

$$F_{2^+} = concat(H_{RSA,2}(U_{2^+}), F_1) \quad m = 3 \tag{14}$$

$$\begin{aligned} F_{(m-1)^+} &= concat\left(H_{RSA,m-1}(U_{(m-1)^+}), F_{(m-2)^+}\right) \\ &\quad m = 4, \ldots, M. \end{aligned} \tag{15}$$

C. Loss Function

This section mainly introduced the hybrid loss function (HLF), which includes the pixel loss, the perceptual loss, and the binary crossentropy (BEC) loss function. Our network is trained through supervised learning with the goal of minimizing the loss function, which can be expressed as

$$\theta = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L_H\left(H_{RSAN}(I^i_{LR}), I^i_{HR}\right) \tag{16}$$

where θ denotes the parameter of the network, N represents the number of input samples, and L_H is the loss function of the RSAN.

Given that realistic degradation processes are difficult to simulate, the network should be subjected to more effective constraints during the training process. The hybrid loss is expressed as

$$L_H = \lambda_1 L_{\text{pix}} + \lambda_2 L_{\text{per}} + \lambda_3 L_{\text{bec}} \tag{17}$$

where the weight coefficients λ_1, λ_2, and λ_3 are used to balance the loss. To calculate the pixelwise difference between the ground-truth and the generated SR image, we use the L_1 function defined as follows:

$$L_{pix} = \|H_{RSAN}(I_{LR}) - I_{HR}\|_1 . \tag{18}$$

Since the VGG [54] network focuses on the deep semantic information, it contributes to the enhancement of network output image clarity, resulting in better visualization. Therefore, in order to fully utilize the feature-level information, we extract features by using a pretrained VGG network, which is used to measure the perceptual loss. In this study, we use Conv5-4 layer in VGG for extracting features. The perceptual loss is calculated as shown as follows:

$$L_{per} = \|f_{\text{VGG54}}(H_{RSAN}(I_{LR})) - f_{\text{VGG54}}(I_{HR})\|_2 \tag{19}$$

where f_{VGG54} is the VGG feature extraction function.

We calculate the BEC loss in the framework of the binary crossentropy loss function, which are shown in (19)

$$\begin{aligned} L_{bec} = & -\Big[I_{HR} \log\left(\hat{H_{RSAN}}(I_{LR})\right) \\ & +(1-I_{HR}) \log\left(1-(\hat{H_{RSAN}}(I_{LR}))\right)\Big] \end{aligned} \tag{20}$$

where L_{bec} represents the BEC loss function as the discriminator. $\hat{H_{RSAN}}(I^{LR})$ denotes the model output after sigmoid activation function, which indicates the probability that the prediction belongs to a ground truth sample. The continuous adjustment of network training is achieved by iterative computation of the above hybrid loss function.

IV. Experiments

A. Experiments Details

We compared the RSAN to seven recent SR methods based on deep learning, which include EDSR [24], DBPN [27], the lightweight residual dense network (RDN) [55], RCAN [28], SAN [29], MHAN [46], SPE [42], and deep unfolding method (LDUM) [56]. We use two publicly available remote sensing datasets for our SR experiments, including remote sensing scene classification (RSCNN7) [57] and UCAS-high resolution aerial object detection dataset (UCAS-AOD). Four prevalent image quality evaluation metrics (i.e., PSNR, SSIM, VIF [58], and ERGAS [59]) are chosen to objectively assess the performance of the model. Furthermore, we conduct experiments on real-scene remote sensing image SR using the Jilin-1 video satellite dataset and introduced the no-reference image quality evaluation metrics image entropy [60] and average gradient [61] to evaluate the performance of the proposed model.

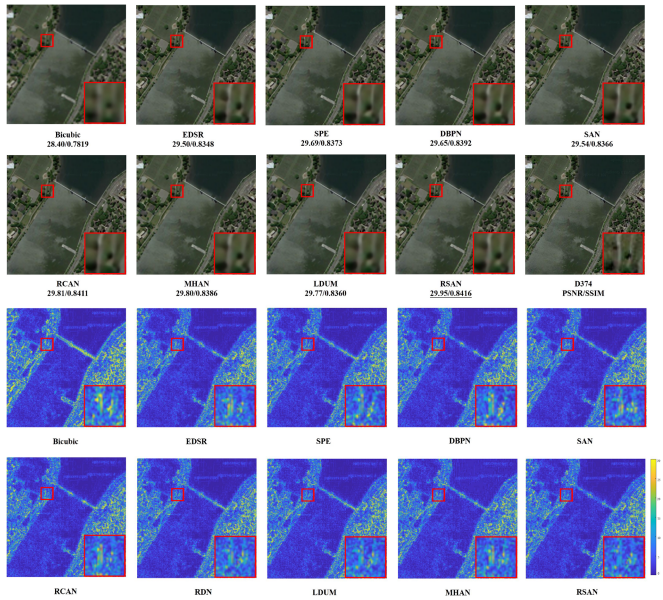

Fig. 3. Visual results with $\times 4$ of the RSSCN7 dataset. The final two rows of images show the error map.

The Adam optimizer is employed for the overall optimization of the network with batch size 16. $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e-8$. The weight of loss function λ_1 is 1, λ_2 is 5.0×10^{-3} and λ_3 is 1×10^{-3}. The learning rate is initialize as $\varepsilon = 1e-4$ and decreases by a factor of 10 for every 500 epochs.

B. Experiments on RSCNN7 Dataset

Wuhan University released the RSSCN7 dataset, which contains a total of 2800 remotely sensed images from seven typical scene categories, with each category containing 400 images. The original image of the RSSCN7 dataset has the pixel size of 400×400. To generate LR images, we apply a downsampling process to the original HR images using a Bicubic interpolation operation with a scale factor and no blur kernel in the MATLAB environment. For the experiment, we select 2100 and 700 images as the training set and test set of RASN, respectively.

Table I compares the quantitative results of the proposed RSAN with recent CNN-based methods under the scale factors of $\times 4$ and $\times 8$ on the RSCNN7 dataset, where the best results are underlined. The experimental results show that the RSAN achieves 0.14 dB than the most competitive general SR method RDN [55] with a scale factor of $\times 4$. Compared to the latest remote sensing image SR methods MHAN [46], SPE [42], and LDUM [56], RSAN achieves a higher PSNR of 0.13/0.08, 0.15/0.19, and 0.11/0.05 dB with the upscaling factor of $\times 4$ and $\times 8$, respectively.

Figs. 3 and 4 show the subjective results on RSCNN7 dataset with scale factors of $\times 4$ and $\times 8$, respectively. The yellow and blue images is represented by the colorbar, which indicates the mean square error (MSE) map between the estimated SR image and the ground truth. The EDSR [24] method are significantly

TABLE I
QUANTITATIVE EVALUATION RESULTS ON THE RSCNN7

Method	Scale	Param/M	PSNR ↑	SSIM ↑	VIF ↑	ERGAS ↓
Bicubic	×4	-	28.43	0.7146	0.3816	2.2483
EDSR [24]	×4	40.09	29.72	0.7718	0.4108	1.9740
DBPN [27]	×4	10.43	29.81	0.7754	0.4157	1.9534
RDN [55]	×4	1.15	29.85	0.7733	0.4194	1.9346
RCAN [28]	×4	15.59	29.77	0.7741	0.4133	1.9628
SAN [29]	×4	15.82	29.76	0.7740	0.4133	1.9640
MHAN [46]	×4	11.35	29.86	0.7768	0.4196	1.9263
SPE [42]	×4	7.84	29.84	0.7754	0.4164	1.9387
LDUM [56]	×4	2.17	29.88	0.7727	0.4204	1.9325
RSAN (Our)	×4	8.21	**29.99**	**0.7786**	**0.4239**	**1.9078**
Bicubic	×8	-	25.77	0.5762	0.2289	3.0669
EDSR [24]	×8	83.22	26.25	0.5954	0.2114	2.9468
DBPN [27]	×8	23.21	26.27	0.6029	0.2192	2.9408
RDN [55]	×8	1.29	26.50	0.6093	0.2388	2.8506
RCAN [28]	×8	15.74	26.41	0.6065	0.2256	2.8850
SAN [29]	×8	15.97	26.41	0.6072	0.2282	2.8843
MHAN [46]	×8	11.51	26.54	0.6140	0.2436	2.8412
SPE [42]	×8	7.99	26.43	0.6080	0.2341	2.8818
LDUM [56]	×8	2.54	26.57	0.6117	0.2453	2.8334
RSAN (Our)	×8	9.72	**26.62**	**0.6165**	**0.2480**	**2.8217**

* The best results are **in bold**.

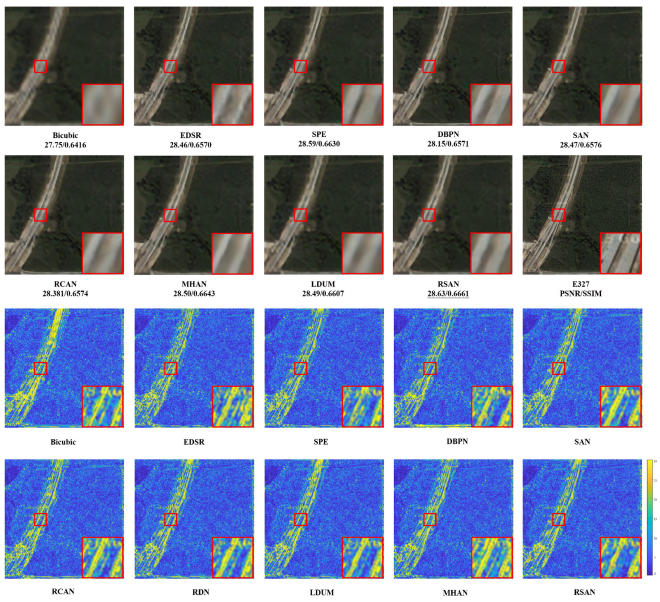

Fig. 4. Visual results with ×8 of the RSSCN7 dataset. The final two rows of images show the error map.

improved compared to the conventional interpolation method (Bicubic). Although the EDSR subjective visual results are fine, the edge information of the generated remote sensing images is significantly insufficient. RCAN [28], SAN [29], and SPE [42] introduce attention mechanisms into the network and obtained better repair results. As the network grows deeper, the number of extracted deep residual features increases, RDN recovers more texture detail on remote sensing images than other single image SR methods.

MHAN also subjoins the attention mechanism, which mixes high-order attention mechanisms with the ability to fully exploit hierarchical features. LDUM utilizes a combination of LR and high-frequency residual images to model HR images, achieving a balance between computational cost and performance. In contrast, our RSAN obtains clearer and better results in saliency regions through a multichannel attention mechanism with multilevel residual feature fusion, which is more faithful to the ground truth. In large-scale factor condition, the RSAN recovers more salient and informative components from LR images and produces more competitive results than other algorithms.

C. Experiments on UCAS-AOD Dataset

The UCAS-AOD dataset is a public satellite image dataset that includes two kinds of targets, automobile and aircraft, and negative background samples. We randomly select 900 of these images with a resolution of 1280 × 689 as the training set. We randomly select 100 HR images and intercept a 200 × 200 pixel portion of them as the test images. To generate LR images, we utilize Bicubic interpolation with a scale factor and no blur kernel in the MATLAB environment to downsample the original HR images.

Table II compares the quantitative results of the proposed RSAN to other CNN-based algorithms with scale factors of ×4 and ×8 on the UCAS-AOD dataset, where the best results are

TABLE II
QUANTITATIVE EVALUATION RESULTS ON THE UCAS-AOD

Method	Scale	Param/M	PSNR ↑	SSIM ↑	VIF ↑	ERGAS ↓
Bicubic	×4	-	30.62	0.7899	0.4490	1.4738
EDSR [24]	×4	40.09	32.71	0.8471	0.5064	1.1786
DBPN [27]	×4	10.43	32.73	0.8480	0.5072	1.1756
RDN [55]	×4	1.15	32.21	0.8341	0.4903	1.2488
RCAN [28]	×4	15.59	32.78	0.8487	0.5091	1.1724
SAN [29]	×4	15.82	32.76	0.8483	0.5083	1.1717
MHAN [46]	×4	11.35	32.65	0.8454	0.5050	1.1894
SPE [42]	×4	7.84	32.51	0.8438	0.5003	1.2003
LDUM [56]	×4	2.17	32.53	0.8427	0.5031	1.2076
RSAN (Our)	×4	8.21	**32.87**	**0.8494**	**0.5128**	**1.1619**
Bicubic	×8	-	25.77	0.5762	0.2289	3.0669
EDSR [24]	×8	83.22	28.72	0.7158	0.3184	1.8114
DBPN [27]	×8	23.21	28.74	0.7189	0.3195	1.8028
RDN [55]	×8	1.29	28.50	0.7073	0.3079	1.8412
RCAN [28]	×8	15.74	28.73	0.7162	0.3119	1.8032
SAN [29]	×8	15.97	28.77	0.7189	0.3170	1.7957
MHAN [46]	×8	11.51	28.56	0.7131	0.3179	1.8447
SPE [42]	×8	7.99	28.59	0.7150	0.3175	1.8408
LDUM [56]	×8	2.54	28.61	0.7127	0.3166	1.8235
RSAN (Our)	×8	9.72	**28.84**	**0.7206**	**0.3247**	**1.7832**

* The best results are **in bold**.

in bold. According to the experimental results, the RSAN has a higher PNSR value of 0.09 dB than the most competitive general SR method RCAN [28] with a scale factor of ×4. Compared with the latest remote sensing image SR method MHAN [46], SPE [42], and LDUM [56], RSAN achieves a higher PSNR of is 0.22/0.28, 0.36/0.25, and 0.34/0.23 dB with the upscaling factor of ×4 and ×8, respectively.

Figs. 5 and 6 show the subjective results on the UCAS-AOD dataset with scale factors of ×4 and ×8, respectively. The last row of images indicate the error map between the estimated SR image and the ground truth. RCAN [28] and SAN [29] can achieve satisfactory outcomes. Nevertheless, their VIF/ERGAS values are lower than the proposed RSAN, particularly at the scale factor of ×8. Since the UCAS-AOD dataset contains only two simple scene categories compared to the RSCNN7 dataset, the proposed RSAN performs significantly better on the UCAS-AOD dataset, especially at large scale factor, compared to the remote sensing image super-resolution methods MHAN, SPE, and LDUM. From the magnified details of the reconstructed images of these methods shown in the images, we observe that the RSAN is capable of obtaining pleasant results, which is reflected in the MSE error maps.

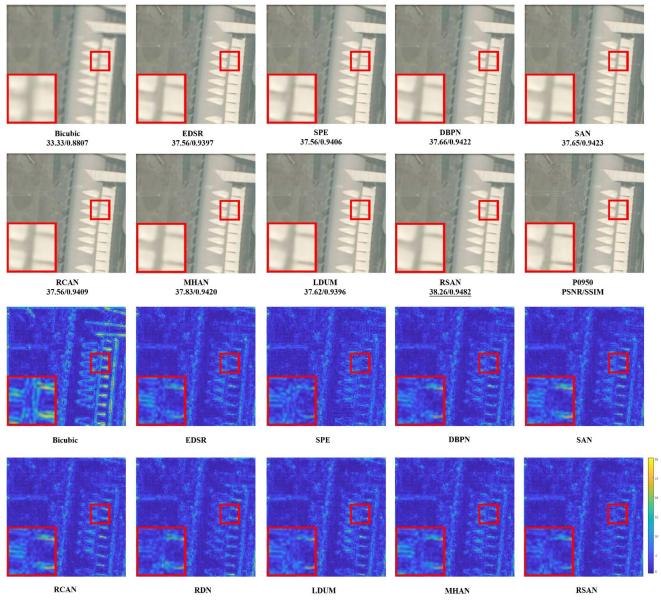

Fig. 5. Visual results with ×4 of the UCAS-AOD dataset. The final two rows of images show the error map.

D. Experiments on Jilin-1 Video Satellite Dataset

In real-world scenarios, the captured satellite images may not meet the precision requirements of many applications due to limitations caused by undersampling and imaging blur of imaging sensors. Under such circumstances, it is essential to utilize SR methods to improve the quality of the LR remote sensing images. To demonstrate the robustness of the proposed RSAN in real-world scenarios, we randomly crop seven remote sensing images of different scenes with a size of 256 × 256 from the Jilin-1 satellite video imageries, and compare with remote sensing SR algorithms through subjective evaluation. As shown

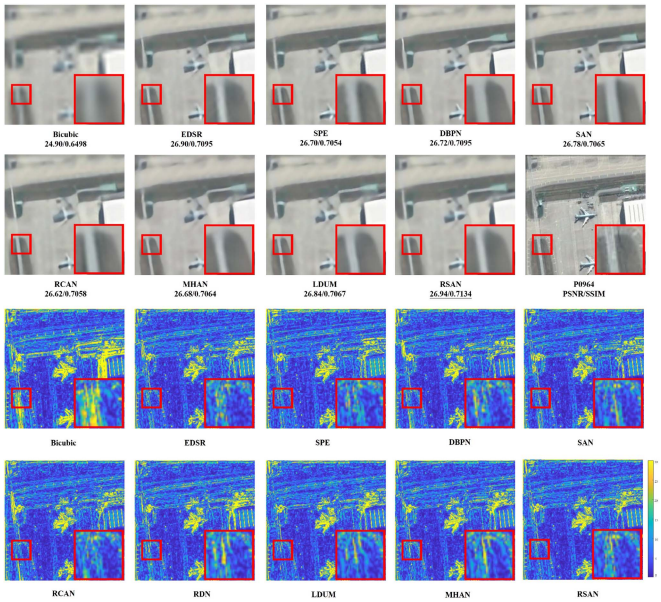

Fig. 6. Visual results with $\times 8$ of the UCAS-AOD dataset. The final two rows of images show the error map.

Fig. 7. Visual and image entropy results of real data with a scale factor of $\times 8$.

in Fig. 7, RSAN has the best reconstruction performance than the other SR method. Specifically, RSAN is able to recover building edges more accurately, while MHAN [46] and SPE [42] exhibit distorted image lines in their results. In the local zoom area, the compared methods produce visible ringing artifacts and blurred outlines, while the proposed RSAN generates sharper edges with fewer jagged lines and artifacts. Based on the above observations, the RSAN can produce visually satisfying high-resolution images with sharp edges and clear boundaries compare to other algorithms.

To further evaluate the SR performance of various methods in practical remote sensing applications, we adopt two no-reference image quality assessment metrics, image entropy (IE) [60] and average gradient (AG) [61]. In SR tasks, IE can be used to measure the complexity of information and texture diversity in an image. Generally, a higher image entropy indicates a larger amount of information and richer texture in the image. The

TABLE III
COMPARISONS RESULT OF IE AND AG ON THE JILIN-1 VIDEO SATELLITE DATASET WITH THE SCALE FACTOR $\times 8$

Method	Bicubic	MHAN	SPE	LDUM	RSAN
IE [60]	4.2244	5.8891	5.9906	6.0519	**6.1862**
AG [61]	5.6801	6.9980	6.9866	7.0182	**7.0310**

* The best results are **in bold**.

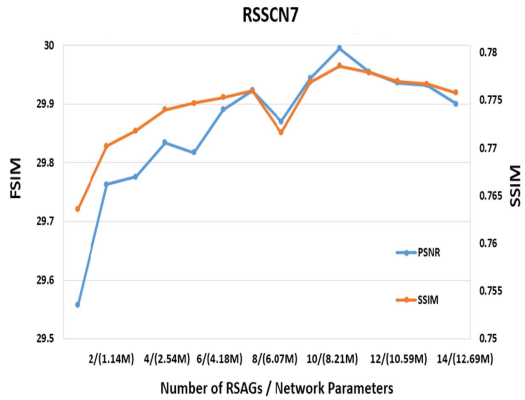

Fig. 8. PSNR/SSIM results of RSAN on the RSSCN7 dataset as m increases from 1 to 14.

average gradient refers to the rate of change in pixel values within an image. Edge and texture details in an image often accompany abrupt changes in pixel values. Therefore, the AG can reflect the level of detail in the edges and textures of the image. The larger the AG and IE, the clearer the image. As shown in Table III, obviously, the proposed RSAN shows a advantage in getting the highest score. In summary, the subjective visual performance and the no-reference image evaluation metrics demonstrate the effectiveness and practicality of the RSAN algorithm in SR remote sensing images.

E. Ablation Analysis

In this section, we first investigate the impact of different numbers of RSAG on the overall network performance, and conduct a series of comparative experiments on the RSSCN7 dataset with a scale factor of 4, as shown in Fig. 8. Subsequently, we introduce ablation studies to verify the effectiveness of the proposed RSAM, RFF, and HLF on the RSSCN7 dataset as shown in Table IV. Furthermore, we visualize the feature maps of the RSAM module to demonstrate that it can help the network focus on regions with rich details. Finally, we compared our proposed method with ResNeSt [52].

As shown in the Fig. 8, The PSNR of RSAN clearly reaches its maximum value when $m = 10$. The value of PSNR increases as m increases until $m = 10$. As m gradually rises to 14, the PSNR of the network gradually decreases by 0.021 dB from its peak value, while the total number of network parameters sharply increases. After carefully considering the trade-offs between network parameters and reconstruction performance, we choose RSAN when the number of RSAGs is 10 as the final network model. In each ablation experiment, we further verify

TABLE IV
ABLATION STUDIES ON THE RSSCN7 DATASET WITH THE SCALE FACTOR ×4

Model	Base	RSAM	RFF	HLF	PSNR↑	SSIM↑
A	✓	-	-	-	29.51	0.7629
B	✓	✓	-	-	29.74	0.7698
C	✓	-	✓	-	29.82	0.7739
E	✓	-	✓	✓	29.85	0.7746
F	✓	✓	-	✓	29.81	0.7747
G	✓	✓	✓	-	29.93	0.7761
H	✓	✓	✓	✓	**29.99**	**0.7786**

* The best results are **in bold**.

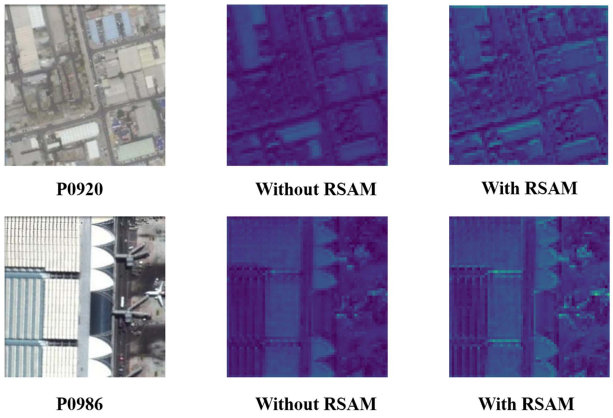

Fig. 9. Comparison of feature map visualization with/without RSAM.

the effectiveness of the proposed final network model RSAN with RSAM, RFF, and HLF. Specifically, the base network is constructed by removing the RFF module from RSAN, replacing RSAM with a depthwise convolutional layer, and using L_1 loss.

Validation on RSAM: We replace the RSAM with the conventional convolutional layers, the SR result showed a decrease about 0.1 dB (see Table IV). The model based on the residual split attention mechanism tends to focus more on the regions rich in detailed information with prominent scenes. In contrast, the conventional convolutional layer treats all feature information in the same way, and direct prediction of high-frequency information tends to produce missing detail information, thus degrading the estimated SR results.

In order to better assess whether the RSAM assists the network in focusing on detail-rich regions, we select images containing samples from two categories, buildings and airplanes, from the UCAS-AOD dataset. For each feature, we visualize and compare the output feature maps of the 5-th RSAG, to demonstrate the effects with and without the RSAM. First, we transfer the channel attention feature maps generated from the fifth RSAM to the CPU for further visualization. Then, we compute the mean of the channel attention feature maps to obtain the mapping results for a single channel. Finally, we utilize the matshow function in the matplotlib library to visualize the channel attention maps as a heatmap. As shown in Fig. 9, the brighter the corresponding region (e.g., building edges and airplane contours) in the visualized feature map, the higher the corresponding value, indicating that the network is more sensitive to these detail information and can better capture the key features of edges and textures in the input data. Therefore, the RSAM can accurately focus on the regions with rich details, thereby improving the network performance.

TABLE V
COMPARISON WITH RESNEST STRUCTURE

Method	Param/M	Times(s)	PSNR ↑	SSIM ↑
ResNeSt [52]	6.02	0.0073	32.64	0.8451
RSAN (Our)	8.21	0.0101	**32.87**	**0.8494**

* The best results are **in bold**.

Validation on RFF: In these ablation studies, we keep the RSAM while removing the both global and local residual feature fusion to verify the proposed residual split attention strategy. It is clear that RSAN performance decreases by more than 0.12 dB when global and local residual feature fusion are eliminated. The optimization of the residual split attention network is guided by aggregating the multilevel global and local residual feature maps of the satellite images, which makes the reconstructed images more accurate. Therefore, without RFF, the reconstructed image will be smooth due to the lack of detail information. This proves that the global and local feature fusion can jointly and adaptively learn hierarchical features in a aggregative way. The artifact removal operation can enhance the edge part of the reconstruction result.

Validation on HLF: To verify the validity of the hybrid loss function, we removed the HLF and set the loss function to L_1 loss. We observed that the SSIM result of the RSAN decreased significantly when the HLF was removed, indicating that the HLF helps to reconstruct the texture and edges of an image in the pixel and perceptual domains, which can improve the estimated SR image to be more approximate to the ground truth image.

Moreover, in order to compare RSAN with the ResNeSt, RSAG is replaced by using the ResNeSt modules and experiments are performed based on the UCAS-AOD dataset with a scale factor of 4. Table V shows the comparison of model size and running time among these methods. In comparison to the ResNeSt, RSAN achieves optimality in PSNR and SSIM. The difference in parameter count between the RSAN and ResNeSt models is attributed to the use of deconvolutional layers in RSAN to learn hierarchical features in a high-resolution potential subspace, which increases the number of parameters. Overall, while RSAN may not outperform other models in terms of running time and number of parameters, it has shown the ability to achieve superior quantitative results.

F. Model Anaysis

In a real remote sensing image SR application scenario, especially in embedded or mobile devices with low computing power, model size and operational efficiency is a key issue. Therefore, we illustrate the comparison of RSAN and other SR networks in terms of the testing time at the scale factor ×4 on Fig. 10.

As shown in Fig. 8, when we set the number of RSAGs to 3 for the simple network, the number of parameters is close to

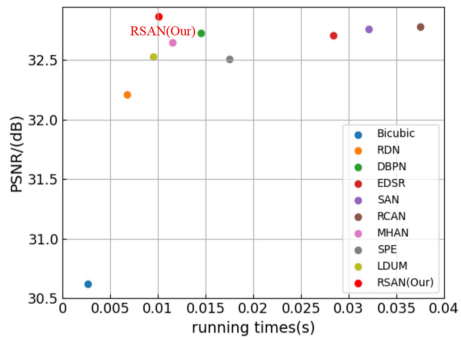

Fig. 10. Running time on the UCAS-AOD dataset with the scale factor ×4.

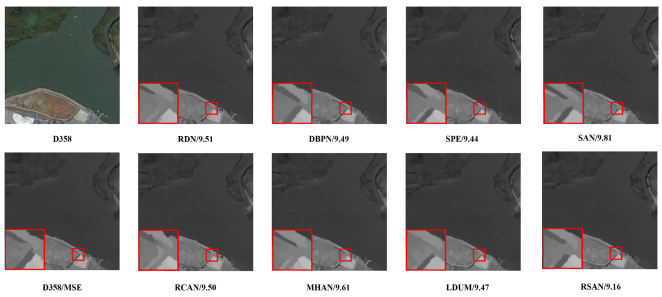

Fig. 11. Segmentation results via the spatial-spectral kernel.

that of the lightweight RDN [55] network, while the algorithm performance is better than that of the noncompact DBPN [27] and EDSR [24] networks. The number of parameters of the complex RSAN (when the number of RSAG is 10) is less than that of the RCAN [28] (15.59 M) and MHAN [46] (11.35 M), which are also based on attention mechanisms, and the image quality assessment results are better. Compared to SPE [42], and LDUM [56], although RSAN does not have an advantage in terms of the number of parameters, it achieves better quantitative results while achieving good inference efficiency, which can provide a suitable network for applications in different scenarios.

G. Performance in Downstream Task

To further validate the effectiveness of the estimated SR images in this article for subsequent image segmentation tasks, we perform unsupervised spatial-spectral kernels [62] as a satellite image semantic segmentation method, and all SR methods use the same parameter settings for image segmentation on the the RSSCN7 dataset.

As shown in Fig. 11, the regions where the proposed RSAN achieves superiority over other SR methods are highlighted in red and green boxes. In the segmentation results obtained by SAN [29], SPE [42], and our proposed RSAN, the buildings along the riverbank (see red box) can be accurately delineated, while other algorithms show varying degrees of misclassification. For the main road, only the method proposed in this article can reconstruct it correctly, which indicates that it outperforms the other compared algorithms and compares favorably with other CNN-based methods. In addition, we used the average MSE value of the three channels of the reconstructed RGB image to measure the direct difference between the SR and ground-truth HR image, quantitatively evaluating the segmentation results. It is clear that the RSAN achieved the best quantitative results.

V. Conclusion

In this article, we propose a novel remote sensing SR method that learns the hierarchical features independently by exploiting the multipath channel feature extraction through the fused multilevel residual features. The proposed method includes four components, i.e., a coarse feature extraction part, the residual split attention groups, a multilevel feature fusion module, and a reconstruction module. We employ the residual split attention group to extract very deep abstract features with long and short skip connection. Meanwhile, the upscale module can remove some of the low-frequency information by performing multiple artifact removal operations, allowing the main network to focus on learning texture and edge information. In addition, to improve the reconstruction capability of RASN, we propose the residual split attention mechanism, which promotes the flow of information in information-rich regions and allows adaptive adjustment of feature weights while maintaining global structural information. Numerous experiments and ablation studies demonstrate the effectiveness of our proposed method, which can achieve superiority over state-of-the-art methods.

References

[1] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4955–4965, Aug. 2014.

[2] J. Li et al., "Feature guide network with context aggregation pyramid for remote sensing image segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9900–9912, 2022.

[3] R. Tsai, "Multiframe image restoration and registration," *Adv. Comput. Vis. Image Process.*, vol. 1, pp. 317–339, 1984.

[4] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.

[5] M. T. Merino and J. Nunez, "Super-resolution of remotely sensed images with variable-pixel linear reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1446–1457, May 2007.

[6] Z. Pan et al., "Super-resolution based on compressive sensing and structural self-similarity for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4864–4876, Sep. 2013.

[7] F. Li, X. Jia, and D. Fraser, "Universal HMT based super resolution for remote sensing images," in *Proc. Int. Conf. Image Process.*, 2008, pp. 333–336.

[8] X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 257–272, Jan. 2013.

[9] F. Li, X. Jia, D. Fraser, and A. Lambert, "Super resolution for remote sensing images based on a universal hidden Markov tree model," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1270–1278, Mar. 2010.

[10] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

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

[12] H. Zhu et al., "Super-resolution reconstruction and its application based on multilevel main structure and detail boosting," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 2065.

[13] 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, 2019, Art. no. 1588.

[14] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2207.

[15] J. Tu, G. Mei, Z. Ma, and F. Piccialli, "SWCGAN: Generative adversarial network combining swin transformer and CNN for remote sensing image super-resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5662–5673, 2022.

[16] Z. Pan, W. Ma, J. Guo, and B. Lei, "Super-resolution of single remote sensing image based on residual dense backprojection networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7918–7933, Oct. 2019.

[17] 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.

[18] T. Lu et al., "Face hallucination via split-attention in split-attention network," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 5501–5509.

[19] Y. Wang, T. Lu, Y. Zhang, and Y. Wu, "Multi-scale self-calibrated network for image light source transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 252–259.

[20] Y. Wang, T. Lu, Y. Zhang, W. Fang, Y. Wu, and Z. Wang, "Cross-task feature alignment for seeing pedestrians in the dark," *Neurocomputing*, vol. 462, pp. 282–293, 2021.

[21] 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.

[22] 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.*, 2016, pp. 1646–1654.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[24] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

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

[26] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 624–632.

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

[28] 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.

[29] 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.*, 2019, pp. 11065–11074.

[30] T. Lu, Y. Wang, J. Wang, W. Liu, and Y. Zhang, "Single image super-resolution via multi-scale information polymerization network," *IEEE Signal Process. Lett.*, vol. 28, pp. 1305–1309, 2021.

[31] L. Wang et al., "Exploring sparsity in image super-resolution for efficient inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4917–4926.

[32] 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.

[33] X. Qin, X. Gao, and K. Yue, "Remote sensing image super-resolution using multi-scale convolutional neural network," in *Proc. 11th U.K.-Eur.-China Workshop Millimeter Waves Terahertz Technol.*, 2018, vol. 1, pp. 1–3.

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

[35] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[36] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2827–2836.

[37] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4133–4141.

[38] 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, 2018, Art. no. 1700.

[39] 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.

[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, 2019, Art. no. 1817.

[41] H. Huan et al., "End-to-end super-resolution for remote-sensing images using an improved multi-scale residual network," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 666.

[42] T. Lu, K. Zhao, Y. Wu, Z. Wang, and Y. Zhang, "Structure-texture parallel embedding for remote sensing image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6516105.

[43] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, and Y. Li, "From artifact removal to super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5627715.

[44] 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.

[45] 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, 2019, Art. no. 2857.

[46] 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.

[47] J. Li, C. Wu, R. Song, Y. Li, and F. Liu, "Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 462–463.

[48] 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.

[49] 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, 2020.

[50] J. Li, Y. Leng, R. Song, W. Liu, Y. Li, and Q. Du, "MFormer: Taming masked transformer for unsupervised spectral reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508412.

[51] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2018–2025.

[52] H. Zhang et al., "RESNest: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2736–2746.

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

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

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

[56] J. Wang, Z. Shao, X. Huang, T. Lu, and R. Zhang, "A deep unfolding method for satellite super resolution," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 933–944, 2022.

[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] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[59] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[60] D.-Y. Tsai, Y. Lee, and E. Matsuyama, "Information entropy measure for evaluation of image quality," *J. Digit. Imag.*, vol. 21, pp. 338–347, 2008.

[61] R. Wang, L. Du, Z. Yu, and W. Wan, "Infrared and visible images fusion using compressed sensing based on average gradient," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2013, pp. 1–4.

[62] 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.

Xitong Chen received the master's degree in software engineering in 2019 from the Wuhan Institute of Technology, Wuhan, China, where he is currently working toward the Ph.D. degree in materials processing engineering with the School of Computer Science and Engineering.

His research interests include image/video processing, computer vision, and artificial intelligence.

Yuntao Wu received the Ph.D. degree in information and communication engineering from the National Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, China, in 2003.

From 2004 to 2006, he was a Postdoctoral Researcher with the Institute of Acoustics, Chinese Academy of Sciences. From 2006 to 2008, he was a Senior Research Fellow with the City University of Hong Kong, Hong Kong. He was a Visiting Researcher with the Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, from 2013 to 2014. He is currently a full-time Professor with the Wuhan Institute of Technology, Wuhan, China, where he is also a Chutian Scholar Project in Hubei Province Distinguish Professor. His research interests include signal detection, parameter estimation in array signal processing, and source localization for wireless sensor networks and biomedicine signal analysis.

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

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

Quan Kong received the B.S. degree in biomedical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2006, and the master's degree in computer science from Kansas State University, Manhattan, KS, USA, in 2017. She is currently working toward the Ph.D. degree in industrial engineering with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China.

She is currently a Teacher with the School of Art and Design, Wuhan Institute of Technology. Her research interests cover deep learning, multimodal image fusion, and interdisciplinary study of generative artificial intelligence in product design.

Jiaming Wang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2022.

He is currently a Teacher with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China. His research interests include image/video processing and computer vision.

Yu Wang received the master's degree in computer technology from the Wuhan Institute of Technology, Wuhan, China, in 2021. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing under the supervision of Prof. Z. Shao with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include image/video processing and computer vision.