GCRDN: Global Context-Driven Residual Dense Network for Remote Sensing Image Superresolution

Jialu Sui, Xianping Ma, *Student Member, IEEE*, Xiaokang Zhang, *Member, IEEE*, and Man-On Pun, *Senior Member, IEEE*

***Abstract*—Superresolution (SR) of remote sensing images aims to restore high-quality information from low-resolution images. Recently, it has witnessed great strides with the rapid development of deep learning (DL) techniques. Despite their good performance, these DL-based models are often ineffective in balancing global and local feature extraction. Moreover, they are usually hindered by the poor image reconstruction capability of the decoder inside their SR models. To cope with this problem, this work proposes a novel global context-driven residual dense network (GCRDN) for satellite image SR based on the encoder and decoder architecture. In particular, the proposed encoder is endowed with nonlocal sparse attention modules incorporated into the residual dense network to learn robust representations from global features. Furthermore, a decoder equipped with back-sampling blocks is devised to fully exploit the feature maps extracted from the encoder. Extensive experimental comparisons based on two multisensor satellite remote sensing datasets confirm that the proposed GCRDN achieves impressive performance in terms of perceptual quality and fidelity.**

***Index Terms*—Convolutional neural network (CNN), nonlocal sparse attention, remote sensing images, superresolution (SR).**

I. Introduction

SUPERRESOLUTION (SR) [1], [2] for remote sensing images aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) observation. It has been widely applied in a wide range of tasks, including military [3], agriculture [4], disaster prevention [5], and natural resource monitoring [6]. Specifically, the HR images generated by SR contain clear spatial texture information of location and essential features of the landscape, and they have been applied in numerous remote sensing vision tasks such as semantic segmentation [7], [8], change detection [9], and object detection [10]. However, it is still challenging to reconstruct fine-grained remote sensing images due to their ultralong-range imaging nature, atmospheric disruptions, and equipment noise [2], [11].

A large number of SR methods have been developed in the literature. Broadly speaking, these existing SR methods can be categorized into three approaches, namely interpolation-based [12], [13], reconstructions-based [14], [15], and learning-based [16], [17], [18], [19] approaches. The interpolation-based approach is an elementary kind of the SR method that generates SR images by increasing the pixel intensities on an up-sampled grid. In contrast, the reconstruction-based SR approach usually relies on explicit prior information to limit the range of potential solutions, assuming that LR images result from HR images after multiple degradations. As a result, the reconstruction-based SR approach is more capable of producing flexible and precise details. Finally, taking advantage of deep learning (DL) techniques, the learning-based approach has demonstrated superior performance by exploiting the statistical correlations between the LR and its matching HR counterpart based on large training sample sets [20]. More specifically, the DL-based SR models are commonly developed upon the encoder–decoder structure in which the encoder extracts representative feature maps from the LR images, whereas the decoder reconstructs the HR images from the feature maps. These DL-based models can be further divided into three categories based on their baseline models, namely convolutional neural network (CNN)-based, generative adversarial network (GAN)-based, and self-attention-based models.

The CNN was first introduced into SR as the baseline model to map an LR image into the HR one in an end-to-end manner [21]. Along the same direction, Kim et al. [22] proposed very deep superresolution (VDSR) by increasing the depth of the network while utilizing the residual learning and gradient clipping to improve the convergence performance of its deep networks. Recently, Lim et al. [23] devised a more comprehensive network named enhanced deep superresolution networks (EDSR) by further enhancing the residual and dense blocks, whereas the authors in [24], [25], [26], and [27] proposed to learn complicated characteristics of ground objects before restoring them into high-quality images in remote sensing. However, the performance of these CNN-based models is handicapped by the limited receptive field of the CNN, incurring insufficient global

Manuscript received 10 February 2023; revised 6 April 2023; accepted 2 May 2023. Date of publication 4 May 2023; date of current version 17 May 2023. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1800800, in part by the Basic Research Project under Grant HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Shenzhen Outstanding Talents Training Fund under Grant 202002, in part by Guangdong Research Projects under Grant 2017ZT07X152 and Grant 2019CX01X104, in part by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence under Grant 2022B1212010001, in part by the National Natural Science Foundation of China under Grant 41801323, and in part by the National Key Research and Development Program of China under Grant 2020YFA0714003. *(Corresponding authors: Xiaokang Zhang; Man-On Pun.)*

Jialu Sui, Xianping Ma, and Man-On Pun are with the School of Science and Engineering, Future Network of Intelligence Institute, Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China (e-mail: jialusui@link.cuhk.edu.cn; xianpingma@link.cuhk.edu.cn; simonpun@cuhk.edu.cn).

Xiaokang Zhang is with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: natezhangxk@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3273081

texture information extraction. On the other hand, GAN-based models have been applied in SR to derive synthetic images of distribution similar to that of the authentic images [28], which results in comparable visual representation [29], [30], [31], [32], [33]. However, these GAN-based models usually suffer from artifacts in the resulting images and loss of high-resolution details of ground objects. Finally, another line of work focuses on the self-attention mechanism for SR, especially the newly proposed transformer structure [34]. The vision transformer [35] has become a popular choice for exploiting long-range contextual information, benefiting from the multihead self-attention mechanism. Furthermore, the Fusformer [36], especially designed for remote sensing images, applied a pure transformer architecture to address the global relationship modeling in feature maps. However, the self-attention-based methods may lead to disturbance from high-frequency noise and loss of detailed information.

In summary, due to the limitation of the receptive field, the CNN-based encoder is ineffective in extracting global features of remote sensing images. In contrast, as the self-attention-based methods focus on calculating the correlation of all elements in the feature maps, they are susceptible to unrelated and noisy contents, thus leading to high computational costs and inaccurate representation. Furthermore, most state-of-the-art methods overlooked the information restoration capability of the decoder, i.e., reconstructing high-quality HR images from abstract features extracted by the encoder.

To remedy these issues, we propose a global context-driven residual dense network (GCRDN) for SR. Specifically, nonlocal sparse attention (NLSA) is first introduced into the residual dense encoder block, which helps exploit the features generated from the residual dense encoder block and effectively aggregate enriched global information. Furthermore, the back-sampling blocks are employed to construct the decoder in which the image reconstruction error is fed back through successive up-sampling and down-sampling blocks. The rich semantic features from the encoder are repeatedly used for the final high-resolution image generation, improving the utilization of semantic information and HR image quality. The main contributions of this article are summarized as follows.

1) A novel nonlocal sparse residual dense (NLRD) encoder is proposed by incorporating NLSA into residual dense networks to capture similar contextual information from a global perspective. The NLRD Encoder is effective and robust as it takes into account the global relationship such as repeated textures in the remote sensing images.
2) The enriched features learned from the encoder are up-sampled and down-sampled successively by the deep back-sampling (DBP) decoder. The circulation of samplings guides the finer reconstruction of the target images while retaining detailed information.
3) Capitalizing on the proposed NLRD encoder and DBP decoder, GCRDN is constructed by leveraging global context extraction and continuous decoding strategy for SR.

The rest of this article is organized as follows: Section II provides a brief overview of different existing DL-based SR models, whereas the architecture of GCRDN is described in Section III. After that, Section IV provides in-depth experimental analyses on GCRDN and other state-of-the-art SR methods. Finally, Section V concludes this article.

II. Related Work

A. CNN-Based SR Framework

The conventional CNN-based encoder–decoder networks have been widely applied in computer vision tasks for decades. Motivated by the great success of classical models such as VGG [37], ResNet [38], DenseNet [39], and MemNet [40] in various practical applications, a number of pioneering SR models have been developed including VDSR [22], RCAN [41], RDN [42], PQA-CNN [43], EDSR [23], MSAN [44], RSI-Net [45], and CTN [46]. These CNN-based SR models are characterized by their convolutional layers for encoding and decoding feature maps. Specifically, RCAN [41] provided a solution for equal treatment of low-frequency and high-frequency features across channels while channel attention was introduced into the residual structure with long skip connections among several residual groups. As a result, the high-frequency information was assigned a higher attention weight than the low-frequency information. This module is also applied in MSAN [44] to perform multilevel feature extraction focusing on the complex structure of remote sensing images. Moreover, PQA-CNN [43] was proposed for SR of remote sensing images by adopting a perceptual quality-assured framework with an uncertainty-driven quantification model to meet the human perceptual requirement. RDN [42] utilized a combination of dense and residual blocks to fully extract the hierarchical features from the LR images. Despite their many advantages, these CNN-based models mainly focus on local information and cannot fully exploit long-range contextual information due to their limited global feature extraction capability. Moreover, another drawback of these CNN-based models is that contextual and spatial information may be lost in the decoding stage, limiting the recovery of the high-resolution information [47].

More recently, GAN has been introduced in image restoration by driving the synthesized results closer to the natural images manifold and discriminating whether it is "real" enough for human perception with the adversarial learning strategy [48]. SRGAN [28], ESRGAN [49], SWCGAN [30], EEGAN [29], MAGAN [31], and CDGAN [32], have been developed for SR. However, their performance is usually hindered by various problems such as model collapse, unstable training, and gradient vanishing during adversarial learning.

B. Self-Attention-Based Enhancement

The transformer was first proposed to model the long-range dependencies in natural language processing before it was introduced to the computer vision field with comparable performance. As the core of the transformer, the self-attention module has been proven to be more effective in arranging long-distance information and attracted much attention. Methods such as SwinIR [50], TranSMS [51], ESRT [52], NLSN [53],

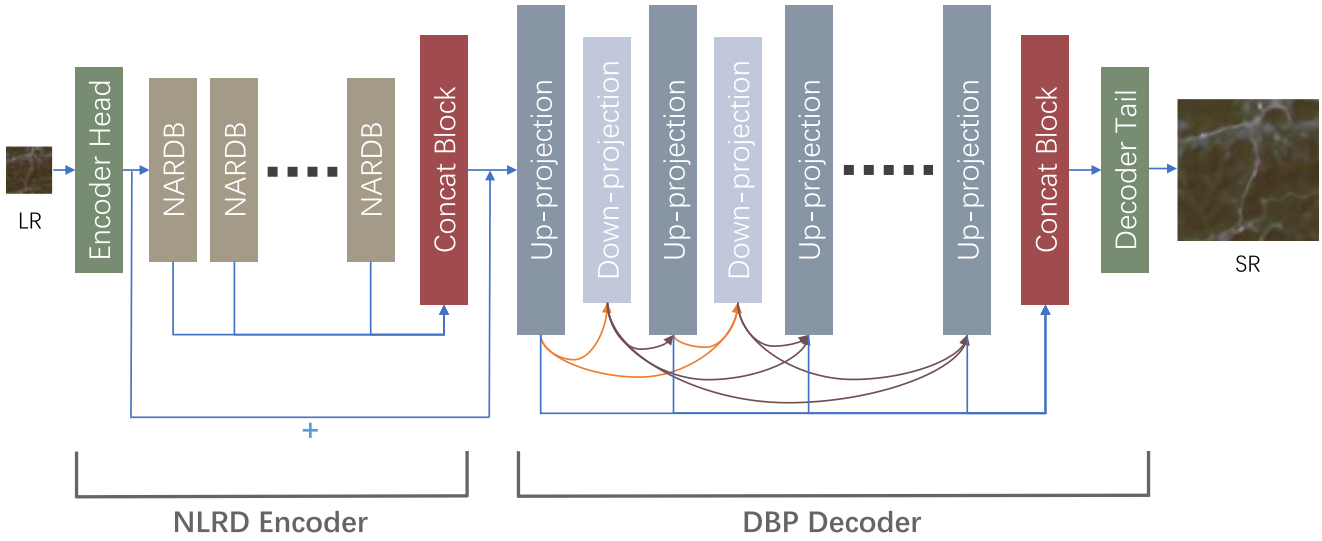

Fig. 1. Framework of the proposed GCRDN. The red and orange lines in the DBP decoder indicate the connections in the down-sampling and up-sampling, respectively. The Concat block concatenates all input matrices before the resulting vector is fed into a convolutional layer to restore the number of feature channels.

Fusformer [36], Interactformer [54], and TransENet [55] are designed based on the self-attention mechanism. In particular, SwinIR [50] adopted a hierarchical transformer with a shifted-window attention mechanism [56]. It consisted of several residual swin transformer blocks to make the interactions between image contents and attention weights. By restricting self-attention computation in non-overlapping local windows, the shifted windowing scheme obtained greater computational efficiency on high-level vision tasks. Moreover, it has been reported that the synergy of CNN and transformer outperforms the pure transformer network. For instance, TranSMS [51] developed a dual-branch encoder built with a vision transformer module and a dense convolutional module. These two branches are used to capture contextual relationships in low-resolution input features and localize high-resolution features, respectively. Furthermore, it has been proposed to embed self-attention modules into the CNN to enhance global information recognition. For instance, NLSN [53] improved the nonlocal sparse attention by identifying the most informative locations that need attentions while ignoring those unrelated regions. Despite that these methods can capture long-range context relationships, they suffer from loss of detailed information in the representation learning and image reconstruction processes. Compared with [53], we combine NSLA with dense residual learning to improve feature expression and introduce the back-sampling strategy to generate finer reconstructed images.

III. Methodology

This section will provide an overview of the proposed method before elaborating on the two essential components of the proposed GCRDN.

A. Network Framework

As shown in Fig. 1, the proposed GCRDN consists of an NLRD encoder and a DBP decoder. In the NLRD encoder, the shallow convolutional features generated by the encoder head are first fed into a series of nonlocal attentive residual dense blocks (NARDBs). These NARDBs are utilized to preserve the feed-forward nature of networks based on a contiguous memory mechanism while extracting local-global contextual features. After that, the multilevel features obtained by the NLRD encoder are concatenated along the channel axis and integrated by convolution operations. Furthermore, the encoded feature maps are fed into the DBP decoder to reconstruct the HR images through the continuous up- and down-sampling processes, in which the encoded features are mapped to the higher resolution feature maps and converted back to the lower resolution repeatedly. Finally, all HR features generated in the up-sampling processes are concatenated and converted into the expected output size before being fed into the decoder tail for SR image reconstruction with the convolution operations.

B. NLRD Encoder

The NLRD encoder aims to learn enriched local-global features of remote sensing images. It consists of an encoder head, several NARDBs and a feature fusion module. The encoder head containing two convolutional layers is used to generate the shallow features of input images. These shallow features are then fed into a series of NARDBs for multilevel representation learning. After that, the feature fusion module composed of convolutional layers is utilized to integrate the multilevel features to enhance representative capabilities.

The architecture of the NARDB is depicted in Fig. 2. As the core component of the NLRD encoder, the NARDB is developed based on hierarchical dense residual learning and the NLSA mechanism.

1) Dense Residual Learning: The dense residual learning in the NARDB can be formulated as follows:

$$O = \text{Conv}([R_J; R_{J-1}; \cdots; R_1; I]) + I \tag{1}$$

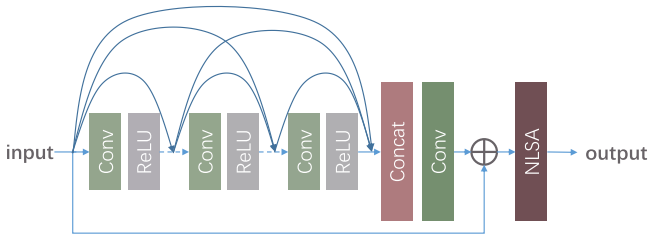

Fig. 2. Architecture of the NARDBs.

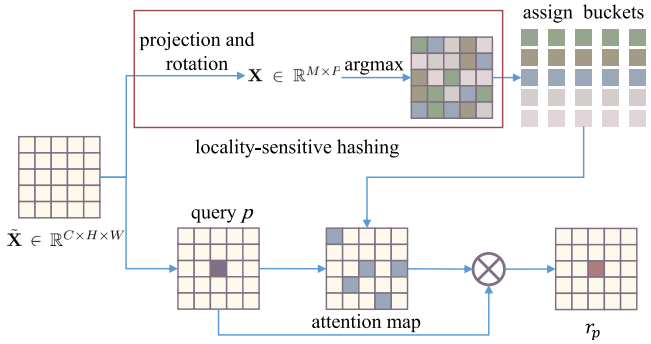

Fig. 3. Process of the NLSA for each feature map.

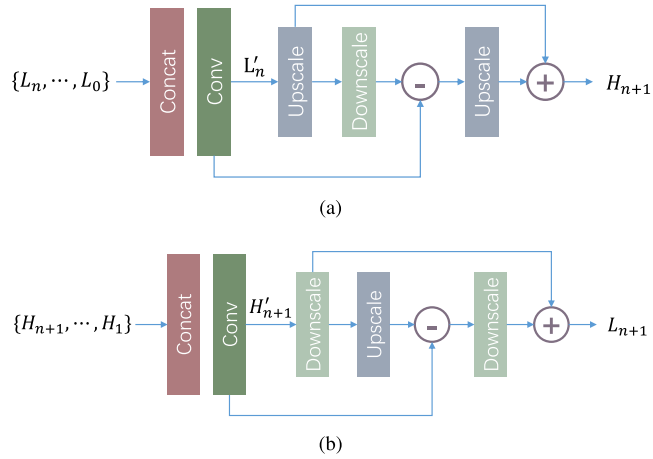

Fig. 4. Architectures of the (a) up-sampling and (b) down-sampling in the DBP decoder.

where O and I are the output and input of the dense residual learning, respectively. Furthermore, $[\cdot]$ stands for the concatenation process, whereas Conv denotes a convolutional operation. In addition, R_j represents the feature maps produced by the jth convolutional layer followed by a ReLU activation function, where $j \in \{1, \ldots, J\}$. Notably, in dense residual learning, the results generated by all convolutional blocks in the previous stage will be connected with the next-stage blocks. After that, an NLSA module is employed to capture global contextual information based on the integrated local features.

2) Nonlocal Sparse Attention (NLSA): Motivated by the observation that remote sensing images typically exhibit repeated patterns, such as vegetation, roads, wasteland, and mountains, NLSA is utilized to capture the global contextual information for better feature extraction. More specifically, NLSA is an improved nonlocal attention operation that partitions the input features into hash buckets to reduce the attention computation.

As shown in Fig. 3, the upper branch presents the first step of NLSA. The feature maps are fed into the spherical locality-sensitive hashing (LSH) algorithm [57] to obtain the attention bucket, while in the bottom branch, for each query p, the attention operation executes in its attention bucket as the calculation range to generate the attention-weighted feature values.

To capture global contextual information from all positions, the input feature map of the NLSA $\tilde{\mathbf{X}} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the feature dimension, the height and width of the feature map, respectively, is first reshaped into a feature sequence $\mathbf{X}' \in \mathbb{R}^{C\times P}$ where $P = H \times W$. Then, the LSH partitions the features $\mathbf{X}'$ into M hash buckets based on the similarity of the angular distances between elements. Input elements with high similarity are mapped into the same bucket. More specifically, LSH first randomly rotates a cross-polytope inscribed into a hypersphere and projects the tensor onto the hypersphere. After that, LSH chooses the closest polytope vertex as a tensor's hash code such that vectors of similar angular distances fall into the same hash bucket. The application of the attention bucket achieves high efficiency and robustness by ignoring other noisy or less-correlated partitions. Denote by $\mathbf{A} \in \mathbb{R}^{M\times C}$ a rotation matrix, the resulting tensor after sampling and rotation is given by

$$\mathbf{X} = \mathbf{A} \cdot \frac{\mathbf{X}'}{||\mathbf{X}'||_2} \tag{2}$$

and subsequently, the hash code at the location p, for $p = 1, 2, \ldots, P$, is defined as

$$\text{hash}(\mathbf{x}_p) = \arg\max_m([\mathbf{x}_p]_m) \tag{3}$$

where $\mathbf{x}_p$ is the pth column of $\mathbf{X}$ and $[\cdot]_m$ stands for the mth entry of the enclosed vector.

As a result, the locations of the same hash code are put into the same bucket. For the feature $\mathbf{x}_p$ at the location p, its bucket index set can be obtained by

$$\mathcal{G}_p = \{q|\text{hash}(\mathbf{x}_q) = \text{hash}(\mathbf{x}_p)\}. \tag{4}$$

Note that $\mathcal{G}_p$ indexes the locations highly related to the location p. Using $\mathcal{G}_p$, we can further compute the NLSA output r_p as follows:

$$r_p = \sum_{q\in\mathcal{G}_p} \alpha_{p,q} \cdot \text{trans}(\mathbf{x}_q) \tag{5}$$

where $\text{trans}(\cdot)$ is a feature transformation function, while $\alpha_{p,q}$ is a weighting coefficient defined as

$$\alpha_{p,q} = \frac{s(\mathbf{x}_p, \mathbf{x}_q)}{\sum_{g\in\mathcal{G}_p} s(\mathbf{x}_p, \mathbf{x}_g)} \tag{6}$$

with $s(\cdot, \cdot)$ being the feature similarity.

Since the input feature map $\mathbf{X} \in \mathbb{R}^{M\times P}$ contains P locations, the set $\mathcal{G}_p$ indexes $|\mathcal{G}_p|$ nonzero elements in $[s(\mathbf{x}_p, \mathbf{x}_1), \ldots, s(\mathbf{x}_p, \mathbf{x}_P)] \in \mathbb{R}^P$, where $|\cdot|$ stands for the cardinality of the enclosed set. Since $\mathcal{G}_p$ contains the pixel locations

TABLE I
QUANTITATIVE EXPERIMENTAL RESULTS ON THE DATASETS

Method	OLI2MSI		Alsat ALL		Alsat agriculture		Alsat urban		Alsat special	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RDN	35.542	0.90491	16.020	0.26747	17.948	0.39181	14.847	0.20473	16.631	0.29596
NLSN	35.503	0.90352	16.063	0.27168	18.050	0.40236	14.877	0.20737	16.673	0.30023
RCAN	35.509	0.90475	16.073	0.27148	17.944	0.39130	**14.921**	0.21089	16.678	0.29903
DBPN	35.109	0.89460	15.901	0.26450	17.331	0.38304	14.822	0.20507	16.530	0.29132
EDRN	34.880	0.88988	16.050	0.26874	17.956	0.38637	14.851	0.20536	16.688	0.29910
EFDN	35.370	0.90108	15.996	0.26519	17.819	0.38247	14.827	0.19987	16.625	0.29725
EDSR	35.471	0.90363	16.030	0.26750	17.757	0.38735	14.874	0.20390	16.667	0.29760
ESRT	34.983	0.89038	16.013	0.26160	17.704	0.35932	14.892	0.20578	16.627	0.28950
TranSMS	35.313	0.89922	16.030	0.26730	17.911	0.37513	14.866	0.20850	16.642	0.29571
SRGAN	35.220	0.89794	15.965	0.26578	17.710	0.38064	14.840	0.20331	16.572	0.29593
ESRGAN	34.668	0.88486	15.972	0.26830	17.609	0.37177	14.889	**0.21346**	16.565	0.29424
Proposed GCRDN	**35.666**	**0.90730**	**16.106**	**0.27595**	**18.178**	**0.40436**	14.894	0.21142	**16.721**	**0.30536**

TABLE II
ABLATION STUDY ON THE OLI2MSI AND ALSAT DATASET FOR NLSA AND BACK-SAMPLING

NLSA	Back-sampling	OLI2MSI		Alsat	
		PSNR	SSIM	PSNR	SSIM
	✓	35.633	0.90690	16.084	0.26965
✓		35.604	0.90601	16.066	0.26780
✓	✓	**35.666**	**0.90730**	**16.106**	**0.27595**

TABLE III
ABLATION STUDY ON THE OLI2MSI AND ALSAT DATASET FOR DENSE CONNECTIONS

Dence connections in encoders	Dence connections in decoders	OLI2MSI		Alsat	
		PSNR	SSIM	PSNR	SSIM
		35.389	0.90138	16.015	0.26690
	✓	35.621	0.90638	16.038	0.27393
✓		35.418	0.90244	16.025	0.26909
✓	✓	**35.666**	**0.90730**	**16.106**	**0.27595**

the query should attend to, the sparsity constraint can be conducted on the NLSA by reducing the number of nonzero entries to the designated chunk size K, i.e., the size of the attention bucket.

Finally, the outputs of all NARDBs are integrated before being fed into the decoder:

$$E_{\text{output}} = \text{Conv}([N_T; N_{T-1}; \cdots; N_1]) + N_0 \tag{7}$$

where E_{output}, N_t, and N_0 are the outputs of the NLRD encoder, the tth NARDB, and the encoder head, respectively, for $t = 1, 2, \ldots, T$. Furthermore, Conv stands for the convolutional operation.

C. DBP Decoder

In sharp contrast to most existing decoders in the encoder–decoder frameworks that directly reconstruct the HR images through progressive convolution and up-sampling in a feed-forward manner, we exploit a DBP decoder to preserve HR components in image reconstructions following an approach similar to [58]. Specifically, the DBP Decoder concentrates on boosting feature sampling at various depths while propagating the reconstruction errors across various stages. As a result, the DBP decoder can learn from different up- and down-sampling operators while retaining the details of HR components. The circulation and interlayer dense connections of up- and down-sampling alleviate the vanishing gradient problem and improve the feature reuse for obtaining better results. Moreover, the up-sampling module takes all down-sampled features as input, while the down-sampling module processes those feature maps produced in each up-sampling unit. In this error feedback strategy, the sampling features in the early stages can guide and constrain the feature expression in the later stages. Without loss of generality, the following discussions will focus on the $(n+1)$th up-sampling and down-sampling operations.

1) Up-Sampling: We denote by $\{L_n, \ldots, L_0\}$ the outputs of the first n down-sampling blocks. The $(n+1)$th up-sampling process is shown in Fig. 4(a) in which the LR feature maps $\{L_n, \ldots, L_0\}$ are first concatenated before being fed into convolution layers. The resulting feature maps denoted as L'_n can be expressed as

$$L'_n = \text{Conv}\left([L_n; L_{n-1}; \cdots; L_1; L_0]\right). \tag{8}$$

L'_n is then first upscaled before being downscaled. After that, the difference between L'_n and its downscaled counterpart is upscaled. Finally, the outputs from both upscale operations are added together to generate the up-sampling output H_{n+1} as follows:

$$H_{n+1} = \text{DC}\left(L'_n - \text{Conv}\left(\text{DC}(L'_n)\right)\right) + \text{DC}(L'_n) \tag{9}$$

where $\text{DC}(\cdot)$ stands for the deconvolution-based upscale operation.

2) Down-Sampling: As presented in Fig. 4(b), the $(n+1)$th down-sampling process is a reverse operation of the up-sampling process. The HR features $\{H_{n+1}, \ldots, H_1\}$ are first concatenated before being fed into convolution layers. The resulting feature maps denoted as H'_{n+1} can be expressed as

$$H'_{n+1} = \text{Conv}([H_{n+1}; H_n; \cdots; H_1]). \tag{10}$$

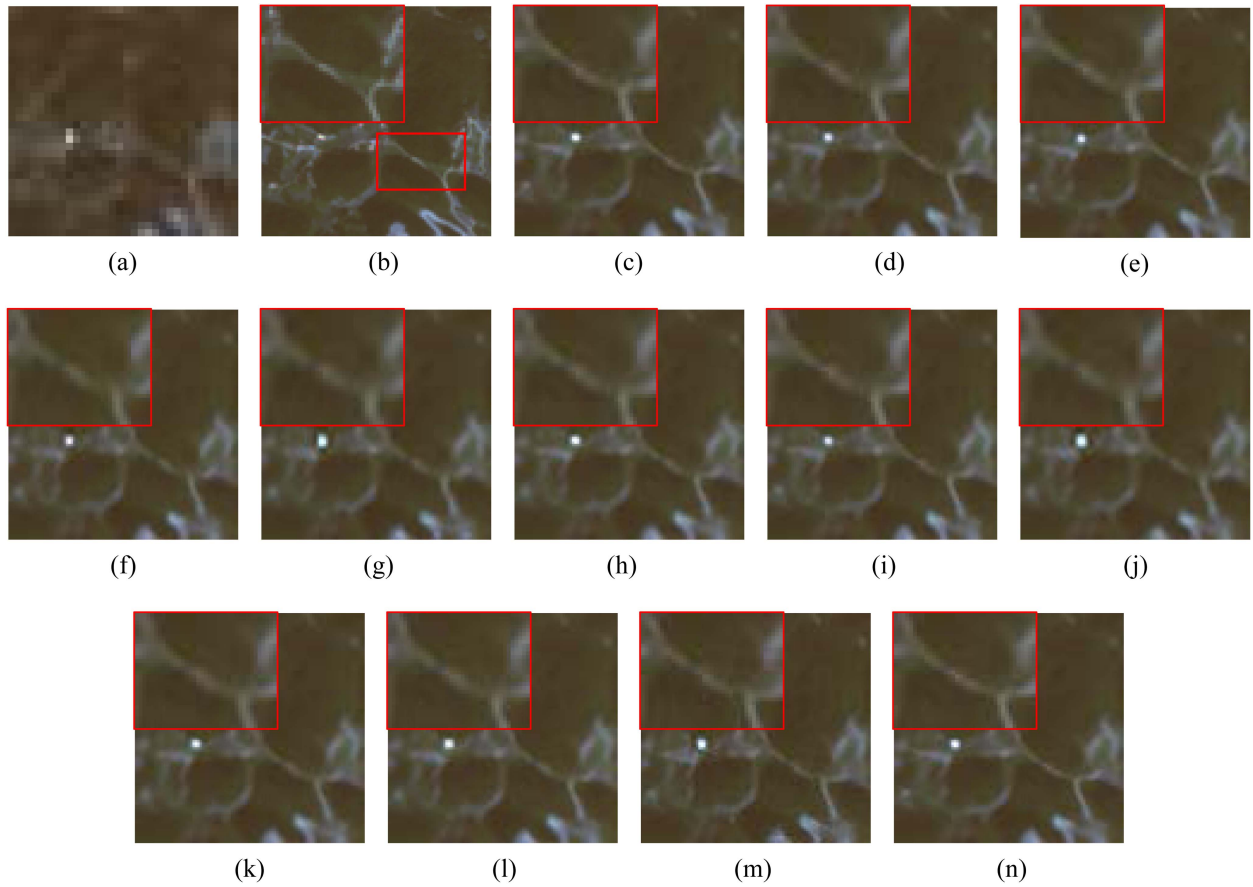

Fig. 5. Visual comparisons on the OLI2MSI dataset. (a) LR. (b) HR. (c) RDN. (d) NLSN. (e) RCAN. (f) DBPN. (g) EDRN. (h) EFDN. (i) EDSR. (j) ESRT. (k) TranSMS. (l) SRGAN. (m) ESRGAN. (n) GCRDN.

TABLE IV
CLASSIFICATION OF THE COMPARED METHODS

Method	Models	Complexity (M)	Memory (MB)	Parameters (M)	Speed (FPS)
CNN-based	EFDN	109.42	2949	7.24	0.7914
	EDSR	716.25	3191	43.68	0.7773
	RCAN	1740.94	6885	106.91	0.7241
	RDN	365.56	5231	22.31	0.7857
	EDRN	29.68	2019	58.53	0.7268
	DBPN	142.28	2949	3.62	0.7946
SAtt-based	ESRT	13.61	2101	0.77	0.7598
	TranSMS	135.12	8413	4.93	0.6889
	NLSN	733.69	6877	44.75	0.7717
	GCRDN	1000.76	13799	29.61	0.7503
GAN-based	SRGAN	14.69	1653	0.73	0.8001
	ESRGAN	9.97	1537	0.62	0.8251

SAtt-based represents the self-attention-based models.

Similarly, H'_{n+1} is then first downscaled before being upscaled. After that, the difference between H'_{n+1} and its upscaled counterpart is downscaled. Finally, the outputs from both downscale operations are added together to generate the $n+1$ down-sampled output L_{n+1} as follows:

$$L_{n+1} = \text{Conv}\left(H'_{n+1} - DC(\text{Conv}\left(H'_{n+1}\right))\right) + \text{Conv}\left(H'_{n+1}\right). \tag{11}$$

Finally, the decoder tail D_{tail} with two convolutional layers gathers all the up-sampled results to compute the final HR images as

$$D = D_{\text{tail}}\left(\text{Conv}([H_U; H_{U-1}; \cdots; H_1])\right) \tag{12}$$

where U is the total number of the up-sampling blocks.

IV. EXPERIMENTS

A. Datasets and Metrics

In this section, two remote sensing datasets, i.e., OLI2MSI [59] and Alsat [60] are employed to evaluate the proposed model. The OLI2MSI dataset comprises Landsat8-OLI and Sentinel2-MSI images with 5225 and 100 pairs of images for training and testing, respectively. Furthermore, Landsat8-OLI images with a spatial resolution of 30 m serve as the LR input and Sentinel2-MSI images with a spatial resolution of 10 m are regarded as the HR ground truth. The Alsat dataset contains 2182 training samples and three subdatasets for testing, namely scenes of "agriculture," "urban," and "special" structures, with 56, 282, and 239 image pairs, respectively. Two widely used metrics, i.e., peak signal to noise ratio (PSNR) and structural similarity (SSIM), are used to quantitatively evaluate the SR performance.

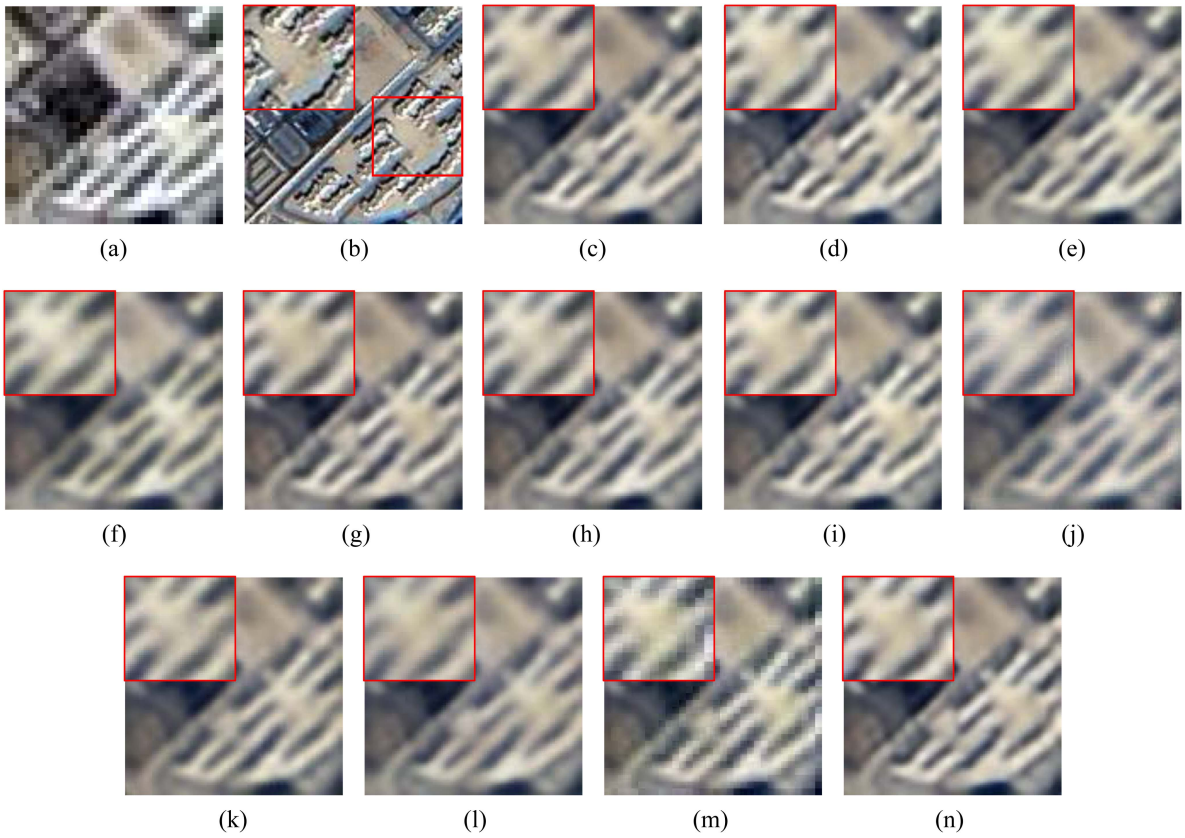

Fig. 6. Visual comparisons on Alsat "urban" set. (a) LR. (b) HR. (c) RDN. (d) NLSN. (e) RCAN. (f) DBPN. (g) EDRN. (h) EFDN. (i) EDSR. (j) ESRT. (k) TranSMS. (l) SRGAN. (m) ESRGAN. (n) GCRDN.

B. Implementation Details

In the proposed network, 16 NARDBs are adopted in the encoder with 3×3 convolution for feature extraction and 1×1 for feature fusion. We use padding operators for all convolutional layers. The chunk size K in the NLSA is set to 144 in OLI2MSI and 25 in Alsat. The total number of up-sampling processes is set to $U = 6$, i.e., six up-sampling operations are utilized in the decoder. The network is trained using the $L1$ loss with a batch size of 16 and a learning rate of 10^{-4}. All images are cropped into patches of 32×32 LR inputs and 96×96 HR outputs in the OLI2MSI dataset. Similarly, 32×32 and 128×128 patches are generated for the training and testing on the Alsat dataset. All experiments are implemented with PyTorch on a single NVIDIA GeForce RTX 3090 GPU with 24 G RAM.

C. Comparisons With Advanced SR Models

To demonstrate the effectiveness of the proposed GCRDN, we compare it against 11 different state-of-the-art models, including RCAN [41], RDN [42], EDSR [23], EFDN [61], DBPN [58], EDRN [47], TranSMS [51], NLSN [53], ESRT [52], SRGAN [28], and ESRGAN [49].

1) OLI2MSI SR: As shown in Table I, the proposed GCRDN achieved the best performance among all the methods under evaluation. In particular, GCRDN equipped with the NLRD encoder and the DBP decoder demonstrated noticeable improvements as compared to our baseline RDN, achieving 0.12 dB and 0.0024 improvement in terms of PSNR and SSIM, respectively. This suggests that the proposed GCRDN is more effective in extracting texture information such as the mountains and roads and generating HR images. Furthermore, inspection of Table I suggests that the proposed GCRDN considerably outperformed the self-attention-based and CNN-based models. This is because that the CNN-based models lacked global features while the self-attention-based models were incapable of effectively utilizing suitable local features. Furthermore, these models also suffered from the weak image reconstruction capability of the decoder. In contrast, the proposed GCRDN benefits from effective feature extraction by leveraging the synergy of the nonlocal attention modules and the back-sampling strategy. Finally, the GAN-based models, e.g., SRGAN and ESRGAN showed worse performances as compared to the CNN-based and self-attention-based models, which indicated that these GAN-based models were less effective for this remote sensing dataset. Fig. 5 shows visual comparisons of images obtained with all methods under evaluation. As presented in Fig. 5, the outline of the roads restored by the proposed GCRDN is more precise and coherent than the others, demonstrating better texture information extraction and reconstruction of the proposed GCRDN.

In summary, the experimental results and visual comparisons discussed previously confirmed that the proposed GCRDN outperformed CNN-based and self-attention models by effectively exploiting global-local information in the encoder and making better feature representation in the decoder.

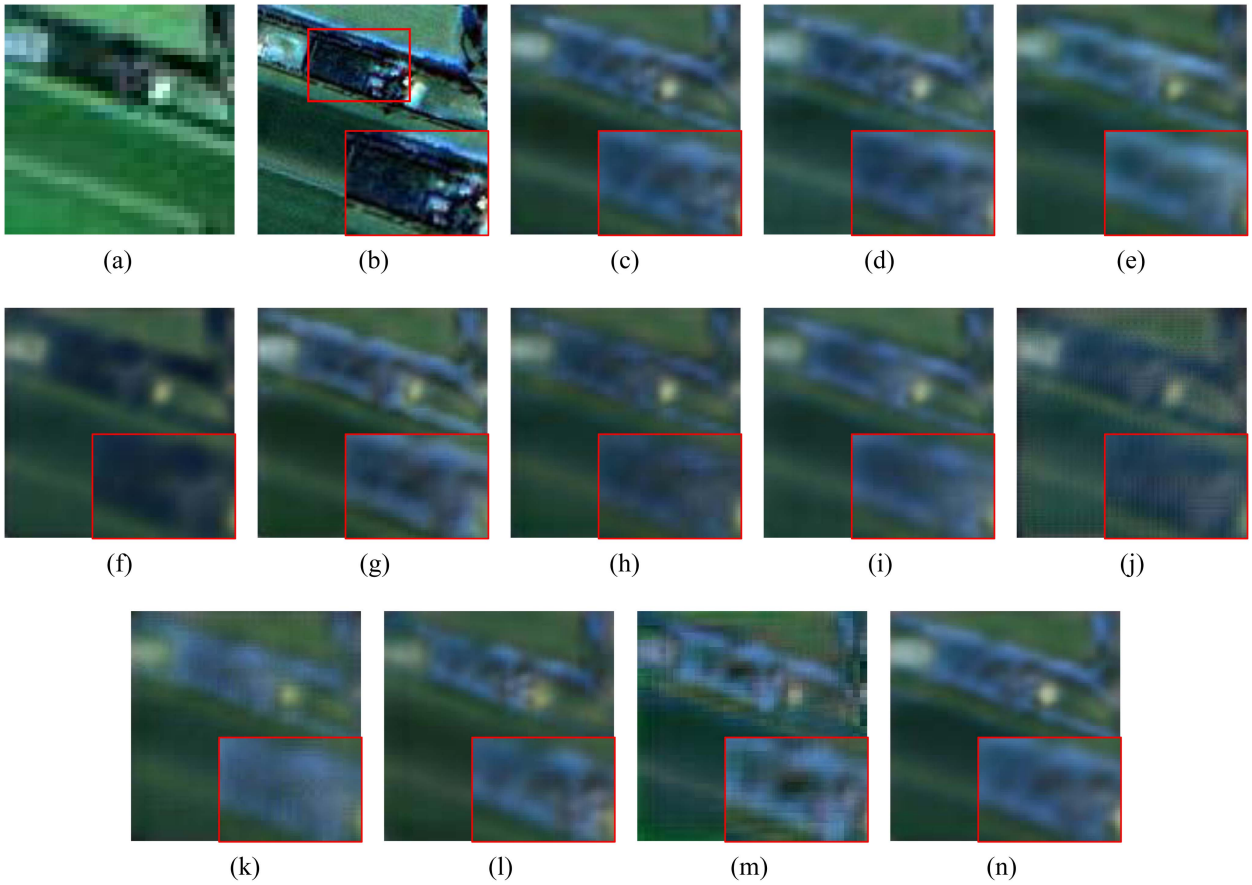

Fig. 7. Visual comparisons on Alsat "agriculture" set. (a) LR. (b) HR. (c) RDN. (d) NLSN. (e) RCAN. (f) DBPN. (g) EDRN. (h) EFDN. (i) EDSR. (j) ESRT. (k) TranSMS. (l) SRGAN. (m) ESRGAN. (n) GCRDN.

2) Alsat SR: For the Alsat dataset, inspection on Table I revealed that the proposed GCRDN achieved the best results among all methods on the collection of three subdatasets. In particular, GCRDN demonstrated 0.06 dB and 0.0079 improvement in terms of PSNR and SSIM, respectively, as compared to our baseline, RDN. Table I further illustrates the performance of all methods in various scenes. It is observed that the proposed GCRDN showed impressive performance in "agriculture," "special," and "urban" scenes. Indeed, the proposed GCRDN achieved the best performance in "agriculture" and "special" test sets because the images of "agriculture" and "special" scenes possess simple and repetitive patterns that can be easily captured and exploited by GCRDN. In "urban" scenes, despite that the performance of GCRDN is worse than that of RCAN in PSNR and that of ESRGAN in SSIM, GCRDN also generated SR images with a high perceptual quality in such a complicated situation. Visual comparisons on the results generated by all methods under consideration in "urban," "agriculture," and "special" scenes are shown in Figs. 6–8, respectively. Clearly, the proposed GRRDN generated clearer SR images with more detailed information and texture. The results from CNN-based methods, especially DBPN, are generally very smooth but blurry. For those GAN-based methods, they have many artifacts, which is highly undesirable in the remote sensing field. In general, SR images provided by the proposed model were visually closest to the HR images while ensuring high fidelity.

D. Ablation Study

Table II presents the ablation investigation on the effect of NLSA and back-sampling. To demonstrate the effect of NLSA and back-sampling, we trained and tested the models by removing the NLSA and replacing the back-sampling decoder with a simple up-sampling decoder on the OLI2MSI and Alsat dataset. The result demonstrates that the NLSA benefited the feature extraction of the residual dense blocks with 0.033 dB and 0.0004 in OLI2MSI and 0.022 dB and 0.0063 in Alsat improvement in terms of PSNR and SSIM, respectively. The enhancement of the decoder resulted in an improvement of 0.062 dB and 0.0013 in OLI2MSI and 0.040 dB and 0.0081 in Alsat in terms of PSNR and SSIM, proving its superior performance. The experiment results confirmed the benefits of the two proposed NLSA and back-sampling in the basic residual dense network.

Furthermore, as shown in Table III, we conduct additional experiments to evaluate the influence of the dense connections in the NARD encoder and DBP Decoder. The dense connections improve the feature extraction and expression performance with 0.029 dB and 0.0050 in the encoder and 0.234 dB and 0.0011 in the decoder in OLI2MSI and Alsat, respectively. Moreover, the

Fig. 8. Visual comparisons on Alsat "special" set. (a) LR. (b) HR. (c) RDN. (d) NLSN. (e) RCAN. (f) DBPN. (g) EDRN. (h) EFDN. (i) EDSR. (j) ESRT. (k) TranSMS. (l) SRGAN. (m) ESRGAN. (n) GCRDN.

last column demonstrates that combining the dense connections in the encoder and decoder could make further improvements.

E. Parameter Analysis

In this section, we investigate the impact of parameter settings on the performance of the proposed GCRDN.

1) Size of Attention Bucket: The sparsity of NLSA is controlled by the size of the attention bucket (chunk size) K. Theoretically, a smaller chunk size can lead to SR images of higher quality under the condition that most correlated elements are identified. We evaluated chunk sizes of $\{4, 25, 100, 144, 225\}$. From Fig. 9(a), it is observed that the PSNR and SSIM performance was insensitive to the chuck size with the largest deviation of 0.007 dB for PSNR and 0.0001 in SSIM across different chunk sizes evaluated, demonstrating the stability of our proposed GCRDN.

2) Stages of Up/Down-Sampling: In this test, we investigated the influence of the stage number of up- and down-sampling operations, i.e., $U - 1$ in the DBP Decoder on feature expression performance. We trained and tested the proposed GCRDN with $\{2, 4, 5, 6, 8\}$ stages of up- and down-samplings. As shown in Fig. 9(b), the SSIM and PSNR performances were not very sensitive to the stage number of up/down-samplings with the largest deviation of 0.061 dB for PSNR and 0.0014 in SSIM across different stage numbers evaluated. Furthermore, four and five stages of up- and down-samplings produced the best SSIM and PSNR, respectively.

F. Visual Activation Maps

Fig. 10 compares the activation maps of the RDN encoder, the RDN decoder, the NLRD encoder, and the DBP decoder using two images. For both images, the activation maps of the NLRD Encoder clearly showed more apparent details than those from RDN, indicating that NLRD achieved effective feature extraction with edge enhancement. Using the same encoder, the DBP decoder demonstrated more effective high-resolution information reconstruction with more evident textures than RDN, especially on those blur boundary lines. The activation maps of the NLRD encoder and the DBP decoder confirmed that the proposed GCRDN could restore high-frequency details while alleviating the blurring artifacts, resulting in sharp and natural edges.

G. Computational Complexity Analysis

Finally, we compare the computational complexity of all methods under evaluation in terms of model complexity, memory, parameters, and inference speed. In particular, we divided the models into three categories, namely the CNN-based, the self-attention-based, and the GAN-based methods. As shown in

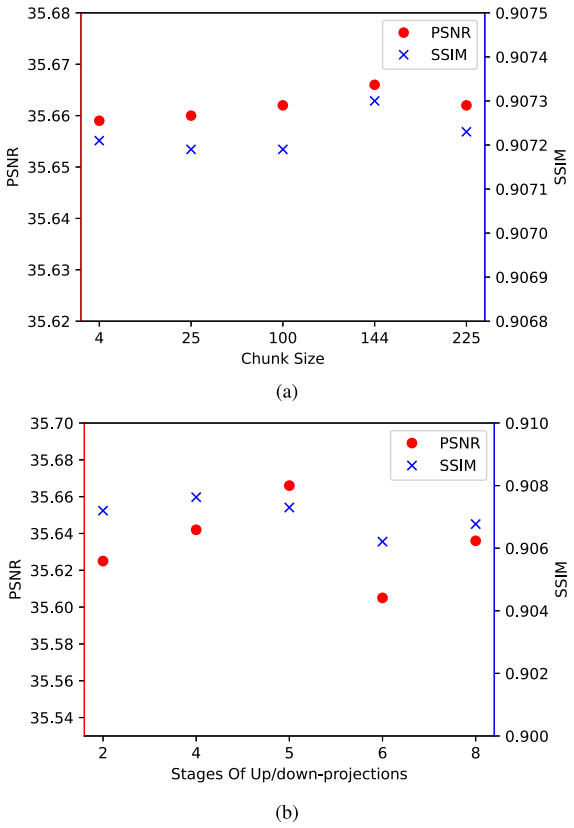

Fig. 9. Experimental comparisons on OLI2MSI with different (a) chunk sizes and (b) numbers of pairs of up-sampling and down-sampling.

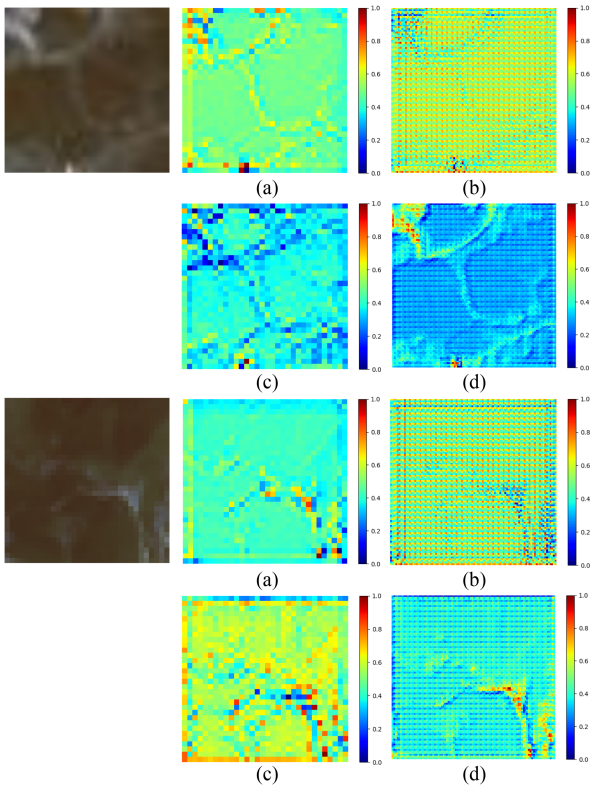

Fig. 10. Visual inspection of the activation maps of (a) RDN encoder, (b) RDN decoder, (c) NLRD encoder, and (d) DBP decoder. Note that the visual activation maps of the DBP decoder were generated with the RDN encoder.

Table IV, the proposed GCRDN has an inference speed comparable to that of other models while requiring higher complexity and memory.

V. Conclusion

In this article, a novel SR method named GCRDN has been proposed by exploiting an NLRD encoder and a DBP decoder to perform effective feature extraction and expression for remote sensing image SR. The NLRD encoder is designed to extract distinct global contextual features whereas the DBP decoder bridges the gap between the enriched features and the final high-quality reconstructed images by the circulation samplings structures. As a result, the proposed GCRDN can effectively characterize the complex content of remote sensing images and restore accurate high-resolution images. Extensive comparative experiments on OLI2MSI and Alsat datasets have confirmed the superior performance of the proposed GCRDN. In the future, we will introduce the diffusion model to further improve the texture reconstruction capability of our model. Moreover, we will extend our model to other image restoration tasks, such as image denoising and cloud removal.

References

[1] H. Chen et al., "Real-world single image super-resolution: A brief review," *Inf. Fusion*, vol. 79, pp. 124–145, 2022.
[2] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: A practical overview," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 314–354, 2017.
[3] R. Hudson and J. W. Hudson, "The military applications of remote sensing by infrared," *Proc. IEEE*, vol. 63, no. 1, pp. 104–128, Jan. 1975.
[4] M. Wójtowicz et al., "Application of remote sensing methods in agriculture," *Commun. Biometry Crop Sci.*, vol. 11, no. 1, pp. 31–50, 2016.
[5] C. Van Westen, "Remote sensing for natural disaster management," *Int. Arch. Photogrammetry Remote Sens.*, vol. 33, no. B7/4; pp. 1609–1617, 2000.
[6] N. Kumar, S. Yamaç, and A. Velmurugan, "Applications of remote sensing and GIS in natural resource management," *J. Andaman Sci. Assoc.*, vol. 20, no. 1, pp. 1–6, 2015.
[7] P. Xu, H. Tang, J. Ge, and L. Feng, "ESPC_NASUnet: An end-to-end super-resolution semantic segmentation network for mapping buildings from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5421–5435, May 12, 2021.
[8] X. Ma, X. Zhang, and M.-O. Pun, "A crossmodal multiscale fusion network for semantic segmentation of remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3463–3474, Apr. 5, 2022.
[9] X. Zhang, W. Yu, and M.-O. Pun, "Multilevel deformable attention-aggregated networks for change detection in bitemporal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 8, 2022, Art. no. 5621518.
[10] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8085–8094, Sep. 13, 2022.
[11] D. Yang, Z. Li, Y. Xia, and Z. Chen, "Remote sensing image super-resolution: Challenges and approaches," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, 2015, pp. 196–200.
[12] P. Thévenaz, T. Blu, and M. Unser, "Image interpolation and resampling," *Handbook Med. Imag., Process. Anal.*, vol. 1, no. 1, pp. 393–420, 2000.
[13] Y. Zhang, Q. Fan, F. Bao, Y. Liu, and C. Zhang, "Single-image super-resolution based on rational fractal interpolation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3782–3797, Aug. 2018.
[14] M. M. Khattab, A. M. Zeki, A. A. Alwan, and A. S. Badawy, "Regularization-based multi-frame super-resolution: A systematic review," *J. King Saud Univ.—Comput. Inf. Sci.*, vol. 32, no. 7, pp. 755–762, 2020.

[15] D. C. Lepcha and B. Goyal, "Medical-modality super-resolution for increased visualisation of intracranial tissue details and structural details," in *Proc. 9th Int. Conf. Rel., Infocom Technol. Optim.*, 2021, pp. 1–9.
[16] Y. Zhang, Y. Zhang, J. Zhang, and Q. Dai, "CCR: Clustering and collaborative representation for fast single image super-resolution," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 405–417, Mar. 2016.
[17] L. Chen, H. Liu, M. Yang, Y. Qian, Z. Xiao, and X. Zhong, "Remote sensing image super-resolution via residual aggregation and split attentional fusion network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9546–9556, 2021.
[18] Z. Shao, L. Wang, Z. Wang, and J. Deng, "Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2663–2674, Aug. 2019.
[19] C. Deng, X. Luo, and W. Wang, "Multiple frame splicing and degradation learning for hyperspectral imagery super-resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8389–8401, Sep. 19, 2022.
[20] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.
[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. 2015.
[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] 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.
[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. 2020.
[25] X. Dong, L. Wang, X. Sun, X. Jia, L. Gao, and B. Zhang, "Remote sensing image super-resolution using second-order multi-scale networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3473–3485, Apr. 2021.
[26] 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.
[27] X. Kang, J. Li, P. Duan, F. Ma, and S. Li, "Multilayer degradation representation-guided blind super-resolution for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 20, 2022, Art. no. 5534612.
[28] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
[29] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.
[30] 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, Jul. 13, 2022.
[31] S. Jia, Z. Wang, Q. Li, X. Jia, and M. Xu, "Multiattention generative adversarial network for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 3, 2022, Art. no. 5624715.
[32] 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.
[33] R. Dong, L. Zhang, and H. Fu, "RRSGAN: Reference-based super-resolution for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 18, 2021, Art. no. 5601117.
[34] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.
[35] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[36] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 27, 2022, Art. no. 6012305.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., SanDiego, CA, USA, May 7–9, 2015.
[38] 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.
[39] 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.
[40] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4539–4547.
[41] 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.
[42] 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.
[43] Y. Zhang et al., "PQA-CNN: Towards perceptual quality assured single-image super-resolution in remote sensing," in *Proc. IEEE/ACM 28th Int. Symp. Qual. Serv.*, 2020, pp. 1–10.
[44] 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.
[45] J. Feng et al., "A deep multitask convolutional neural network for remote sensing image super-resolution and colorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 24, 2022, Art. no. 5407915.
[46] S. Wang, T. Zhou, Y. Lu, and H. Di, "Contextual transformation network for lightweight remote-sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 1, 2021, Art. no. 5615313.
[47] G. Cheng, A. Matsune, Q. Li, L. Zhu, H. Zang, and S. Zhan, "Encoder-decoder residual network for real super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1–10.
[48] X. Zhang, X. Zhu, X. Zhang, N. Zhang, P. Li, and L. Wang, "SegGAN: Semantic segmentation with generative adversarial network," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data*, 2018, pp. 1–5.
[49] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 1–23.
[50] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1833–1844.
[51] A. Güngör, B. Askin, D. A. Soydan, E. U. Saritas, C. B. Top, and T. Çukur, "TranSMS: Transformers for super-resolution calibration in magnetic particle imaging," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3562–3574, Dec. 2022.
[52] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 457–466.
[53] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 15, 2022, pp. 3517–3526.
[54] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531715.
[55] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 16, 2021, Art. no. 5615611.
[56] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[57] K. Terasawa and Y. Tanaka, "Spherical LSH for approximate nearest neighbor search on unit hypersphere," in *Proc. 10th Int. Workshop Algorithms Data Struct.*, 2007, pp. 27–38.
[58] 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.
[59] J. Wang et al., "Multisensor remote sensing imagery super-resolution with conditional GAN," *J. Remote Sens.*, vol. 2021, pp. 1–11, 2021.
[60] A. Djerida, K. Djerriri, M. S. Karoui, and M. El Amin, "A new public Alsat-2B dataset for single-image super-resolution," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 8095–8098.

[61] Y. Wang, "Edge-enhanced feature distillation network for efficient super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 777–785.

Jialu Sui received the bachelor's degree in computer science and technology from Shandong University, Weihai, China, in 2021. She is currently working toward the M.Phil. degree in computer and information engineering with the Chinese University of Hong Kong, Shenzhen, China.

Her research interests include remote sensing and deep learning.

Xianping Ma (Student Member, IEEE) received the bachelor degree in geographical information science from Wuhan University, Wuhan, China, in 2019. He is currently working toward the Ph.D. degree in computer and information engineering with the Chinese University of Hong Kong, Shenzhen, China.

His research interests include remote sensing image analysis and deep learning.

Xiaokang Zhang (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018.

From 2019 to 2022, he was a Postdoctoral Research Associate with The Hong Kong Polytechnic University, Hong Kong, and The Chinese University of Hong Kong, Shenzhen, Shenzhen, China. He is currently a specially-appointed Professor with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan. He has authored or co-authored more than 20 scientific publications in international journals and conferences. His research interests include remote sensing, computer vision, and deep learning.

Dr. Zhang serves as a Reviewer for more than ten renowned international journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and *Information Fusion*.

Man-On Pun (Senior Member, IEEE) received the B.Eng. degree in electronic engineering from The Chinese University of Hong Kong, Shenzhen (CUHKSZ), Shenzhen, China, in 1996, the M.Eng. degree in computer science from the University of Tsukuba, Tsukuba, Japan, in 1999, and the Ph.D. degree in electrical engineering from the University of Southern California at Los Angeles, Los Angeles, CA, USA, in 2006.

He was a Postdoctoral Research Associate with Princeton University, Princeton, NJ, USA, from 2006 to 2008. He held research positions with Huawei, Milford, NJ, USA, the Mitsubishi Electric Research Labs, Boston, MA, USA, and Sony, Tokyo, Japan. He is currently an Associate Professor with the School of Science and Engineering, CUHKSZ. His research interests include artificial intelligence, Internet of Things, and applications of machine learning in communications and satellite remote sensing.

Dr. Pun was the recipient of best paper awards from the IEEE Vehicular Technology Conference 2006 Fall, the IEEE International Conference on Communication 2008, and the IEEE Conference on Computer Communications 2009. He is the Founding Chair of the IEEE Joint Signal Processing Society-Communications Society Chapter, Shenzhen. He served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2010 to 2014.