Context Aware Edge-Enhanced GAN for Remote Sensing Image Super-Resolution

Zhihan Ren, Lijun He, and Jichuan Lu

***Abstract*—Remote sensing images are essential in many fields, such as land cover classification and building extraction. The huge difference between the directly acquired remote sensing images and the actual scene, due to the complex degradation process and hardware limitations, seriously affects the performance achieved by the same classification or segmentation model. Therefore, using super-resolution (SR) algorithms to improve image quality and achieve better results is an effective method. However, current SR methods only focus on the similarity of pixel values between SR and high-resolution (HR) images without considering perceptual similarities, which usually leads to the problem of oversmoothed and blurred edge details. Moreover, there is little attention to human visual habits and machine vision applications for remote sensing images. In this work, we propose the context aware edge-enhanced generative adversarial network (CEEGAN) SR framework to reconstruct visually pleasing images that can be practically applied in actual scenarios. In the generator of CEEGAN, we build an edge feature enhanced module (EFEM) to enhance the edges by combining the edge features with context information. Edge restoration block is designed to fuse multiscale edge features enhanced by EFEM and reconstruct a refined edge map. Furthermore, we designed an edge loss function to constrain the generated SR and HR similarity at the edge domain. Experimental results show that our proposed method can obtain SR images with a better reconstruction performance. Meanwhile, CEEGAN can achieve the best results on classification and semantic segmentation datasets for machine vision applications.**

***Index Terms*—Edge enhancement, generative adversarial network (GAN), remote sensing images, super-resolution (SR).**

I. Introduction

SUPER-RESOLUTION (SR) is the process of restoring a high-resolution (HR) image from a given low-resolution (LR) image. With the development of remote sensing image application technology, remote sensing images are extensively used in hyperspectral application [1], [2], [3], [4], [5], [6], [7], [8], object detection [9], [10], change detection [11], [12], [13], and other fields. However, image SR is an ill-posed problem because one LR image may degenerate from several different HR images. The SR algorithm is to find an optimal HR image from all possible solutions. The quality of remote sensing images is limited by hardware equipment and natural environment interference, such as clouds and fog, which leads to serious degradation of the acquired remote sensing images that do not match the actual scene seriously. However, the natural image SR algorithms neglect these factors, leading to unsatisfactory results in remote sensing images. Specifically, incorrect edges can cause semantic segmentation models to be unable to distinguish pixels at the edge or image classification models to be unable to correctly classify images. Therefore, the development of a remote sensing image SR algorithm that can accommodate both human visual perception and machine vision applications is crucial.

Over the past few decades, SR has attracted great attention from researchers and many SR methods have emerged. Existing methods can be divided into two main categories: 1) traditional methods and 2) deep learning-based methods. Traditional methods can be further categorized into interpolation methods and reconstruction methods. Interpolation methods use the same kernel without considering the position of the pixel point. Interpolation methods, such as nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation [14], can improve the image resolution based on the content of the image itself but cannot provide more information. However, they may cause edge blurring and fail to achieve the desired visual quality. The reconstruction-based methods [15], [16] solve the problem from the perspective of image degradation models, assuming that the HR images are obtained from the LR images with appropriate motion transformation, blurring, and noise. These methods constrain the generation of SR images by extracting information from LR images and combining it with prior knowledge. However, reconstruction-based methods suffer from the problems of complex optimization methods and high computational costs.

With the development of deep learning, SR algorithms based on convolutional neural network (CNN) have been proposed [17]. Based on this, an amount of research [18], [19], [20], [21], [22], [23], [24], [25] has been devoted to optimizing networks by using L1 or L2 loss functions. These methods

Manuscript received 12 October 2023; revised 7 November 2023; accepted 10 November 2023. Date of publication 15 November 2023; date of current version 14 December 2023. This work was supported in part by the National Key R&D Program of China under Grant 2022ZD0115802, in part by the Central Guidance on Local Science and the Technology Development Fund under Grant 2022ZY1-CGZY-01HZ01, in part by the Natural Science Foundation of Sichuan Province under Grant 2022NSFSC0966, and in part by the Key R&D Project in Shaanxi Province under Grant 2023-ZDLNY-65. *(Corresponding author: Lijun He.)*

Zhihan Ren is with the Shaanxi Key Laboratory of Deep Space Exploration Intelligent Information Technology, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: renzh0729@gmail.com).

Lijun He is with the Shaanxi Key Laboratory of Deep Space Exploration Intelligent Information Technology, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Sichuan Digital Economy Industry Development Research Institute, Chengdu, Sichuan 610036, China (e-mail: lijunhe@mail.xjtu.edu.cn).

Jichuan Lu is with the China Mobile Communications Group Shaanxi Company Ltd., Xi'an 710077, China (e-mail: lujichuan@sn.chinamobile.com).

Digital Object Identifier 10.1109/JSTARS.2023.3333271

use CNN to learn the implicit mapping between pairs of LR and HR images and then predict the HR images corresponding to the LR images based on the learned mapping relationship. However, these methods select PSNR as the evaluation metric and generate images with high PSNR values that may not align with human visual perception. In addition to the PSNR-oriented method, there is another type of methods [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36] that introduce generative adversarial network (GAN) [37] into the field of SR, which aim to generate more realistic images. Although the images generated by GAN-based methods are more realistic compared to PSNR-oriented methods, the high-frequency details learned by GAN-based methods may not be consistent with the actual situation due to the lack of constraints and the severe degradation of remote sensing images. Jiang et al. [29] noticed that existing methods cannot generate correct edges and proposed a denoising approach through mask processing. However, this method solely focused on extracting and utilizing features from the edge domain, neglecting the valuable contextual semantic information present in remote sensing images. Moreover, it does not impose any constraints on the enhanced edges, resulting in suboptimal performance. Recently, the transformer and diffusion model has been widely used in the field of computer vision [38], [39], [40]. Some works have extended the transformer and diffusion model to the field of SR. SwinIR [41] proposed an image reconstruction model based on Swin Transformer [40]. Transformer-based enhancement network (TransENet) [42] proposed a multistage enhancement structure based on transformer. Inspired by the denoising diffusion probabilistic model, Image super-resolution via iterative refinement (SR3) [43] performed SR images through a random iterative denoising process.

In general, there are still some limitations of the existing algorithms.

1) The generated images by the PSNR-oriented method may not align with human visual perception since they ignore the perceptual similarity, which results in oversmoothed images.
2) The practical applications of GAN-based methods may be limited due to their susceptibility to noise, and inaccuracies in generating essential high-frequency information in images, especially in the case of remote sensing images with complex degradation processes.

To address the above problems, we propose an SR framework for joint context semantic information for edge enhancement of remote sensing images, named context aware edge-enhanced generative adversarial network (CEEGAN). The main contributions of this work are summarized as follows.

1) *An SR framework for both human vision and machine analysis:* The existing SR framework is primarily focused on restoration at the pixel level, and a high PSNR may not satisfy the requirements for high-level computer vision tasks in practical scenarios. Remote sensing images usually contain a multitude of objects with different scales, shapes, and complex spatial relationships. Therefore, remote sensing images possess highly rich contextual semantic features. However, in the complex degradation process of remote sensing images, the contextual semantic features can be severely affected. CEEGAN performs the information exchange between contextual semantic features and edge features, and integrates contextual features into the image, aiming to restore the texture information as much as possible in the generated SR images. Both types of features complement each other, to satisfy the needs of both human and machine vision.
2) *An edge enhancement module based on context-guidance:* To enhance the images with blurred edge details, we explore and integrate multiscale (MS) edge features and context semantic features in the edge feature enhanced module (EFEM). The high-level context semantic features are beneficial to guide the extraction and enhancement of edge features. Therefore, we keep the information interaction between the context semantic and edge branches to help the edge branch understand the high-level semantic information in the image. Moreover, we fuse the MS features enhanced by EFEMs and reconstruct the edge map through the edge restoration block (ERB).
3) *An edge loss function to generate more realistic image:* Due to the neglect of edges by L1 and L2 loss, which leads to the oversmooth images, and the sensitivity of GAN-loss to noise, we expand the attention of SR models from the pixel domain to the edge domain by designing edge loss function that constrains the SR and HR image in the edge domain, making the generated images more realistic and preserving important edge details.

The rest of this article is organized as follows. Section II summarizes the related works. Section III introduces the details of CEEGAN. The experimental results are given in Section IV. Finally, Section V concludes this article.

II. Related Work

A. PSNR-Oriented SR Models

Since CNN was introduced to SR by Dong et al. [17], a super-resolution convolutional neural network (SRCNN) with three convolutional layers was pioneered. SRCNN utilized the L2 loss function to optimize the PSNR, aiming to improve its performance beyond traditional methods with the strong nonlinear fitting capability of CNN. With the introduction of ResNet [44], SRResNet [26] introduced the residual network in the SR network to combine the low-level feature and deep-level feature to improve the network learning ability. To enhance the performance of the network, Lim et al. [27] proposed the enhanced deep residual network (EDSR), which removed the BN layer and increased the model size without increasing computational resources. Furthermore, Zhang et al. [18] proposed a residual channel attention network (RCAN) with a channel attention mechanism, which can adaptively rescale channel features. With the development of the transformer, Chen et al. [45] constructed the first transformer model for image SR, and Liang et al. [41] proposed SwinIR, an image reconstruction model based on the swin transformer [40].

Remote sensing images are more complex than natural images in terms of the degradation process and contain objects of varying sizes and shapes, which makes SR of remote sensing

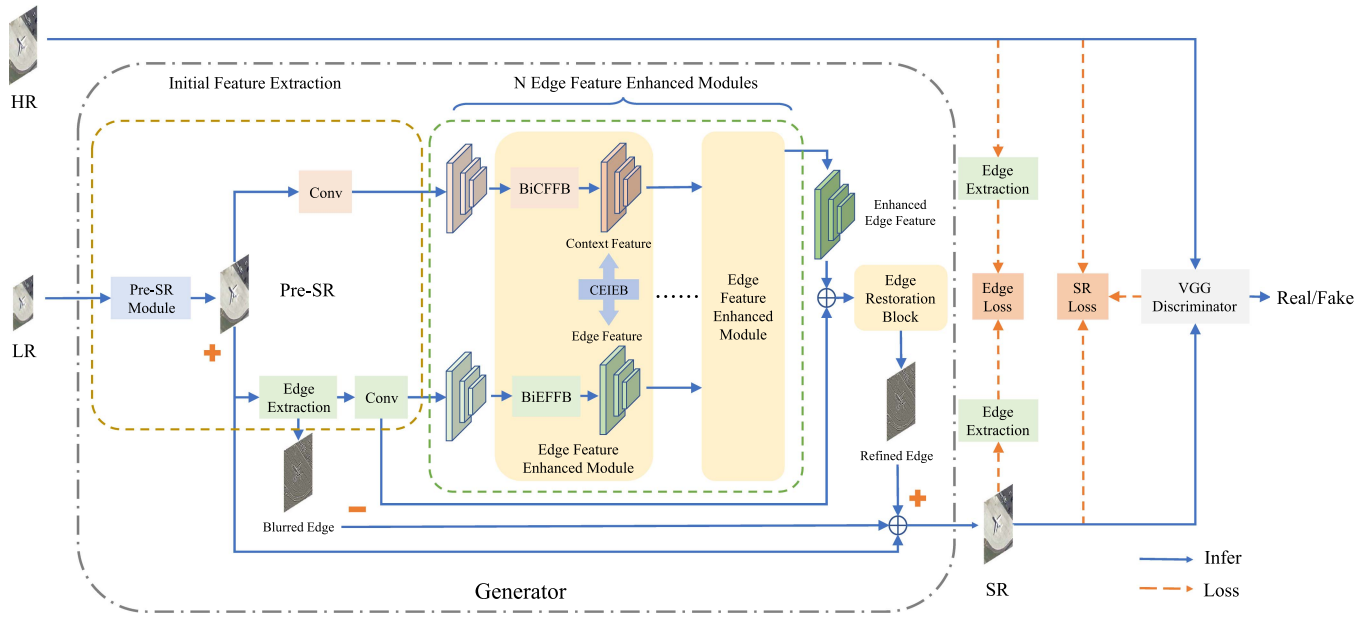

Fig. 1. Overall structure of the CEEGAN. Edge extraction denotes the Laplacian operator. The blue line indicates the network flow of CEEGAN. The orange line indicates the calculation process of the loss function of CEEGAN. SRLoss contains L1 loss, adversarial loss, and perceptual loss.

images more challenging. Therefore, some works proposed SR algorithms for remote sensing images, considering their unique characteristics and challenges. Lei et al. [19] designed a local-global combined network, which combined local and global information by cascading shallow and deep feature mappings. To further integrate information from different depths, Zhang et al. [20] proposed the mixed high-order attention network, which made full use of hierarchical features through high-order attention. Some works addressed the problem of remote sensing image SR from the direction of MS features. The MS attention network [21] employed convolutional with different kernel sizes to extract MS features of remote sensing images and uses the channel attention mechanism to fuse features at different scales. Dong et al. [22] proposed a second-order MS super-resolution network to maximize the use of learned MS information by exploiting small-difference and large-difference features at the local and global levels, respectively. Hybrid-scale self-similarity exploitation network (HSENet) [23] used self-similarity to learn internal recurrence models in single-scale and cross-scale in remote sensing images, achieving stronger feature representation. In contrast to the abovementioned methods, TransENet [42] proposed a transformer-based MS enhancement structure to MS high and low-dimensional features.

B. GAN-Based SR Models

Ledig et al. [26] first applied GAN to image SR reconstruction and proposed SRGAN, which increased perceptual loss and adversarial loss to make the generated images more realistic. To fully leverage the features across different layers, ultradense GAN (udGAN) [36] proposed the ultradense residual block, which reforms the internal layout of the residual block into a two-dimensional matrix topology. To improve the judgment of discriminators in GAN, enhanced SRGAN (ESRGAN) [28] proposed relativistic GAN that allowed the discriminator to predict relative realness instead of the absolute value and used the network interpolation method to balance the conflict between objective and subjective evaluation metrics. Similarly, coupled-discriminated GANs [46] proposed a discriminator that makes judgments based on SR images and HR images to better distinguish the input. To solve the problem of unclear image generation edges, edge-enhanced GAN (EEGAN) [29] proposed an edge enhancement strategy by purifying noisy images to generate precise edges and enhance image contours. Multiattention GAN [30] proposed branch attention to integrating upsampled LR images with high-level features. Different from the PSNR-oriented SR methods, which pursue high PSNR value, the GAN-based methods prefer to generate more realistic images. However, the GAN-based methods are sensitive to noise, which leads the generator to add incorrect high-frequency details to SR images. This shortcoming limits the performance of SR images in subsequent high-level computer vision tasks.

III. Method

A. Overview of the Proposed CEEGAN

Fig. 1 illustrates the overall framework of CEEGAN. The generator of CEEGAN can be divided into the following three parts: Initial feature extraction (IFE), EFEM, and ERB.

The original LR image $I_{\mathrm{LR}} \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width of the image, respectively, is first processed by presuper-resolution module (PSRM) to generate a Pre-SR image $I_{\mathcal{P}} \in \mathbb{R}^{4H\times 4W\times 3}$, which can achieve reconstruction of most regions in LR images except for the edges. Pre-SR image is simultaneously fed into the edge and context semantic feature extraction layer to obtain corresponding features. In the

EFEM, the bidirectional edge feature fusion block (BiEFFB) and bidirectional context feature fusion block (BiCFFB) are used for bidirectional weighted fusion to learn the importance of features between different scales. Simultaneously, we designed the context edge information exchange block (CEIEB) to exchange information between the two branches, which is helpful in guiding edge information using advanced context semantic information. For the MS enhanced features of EFEM output, ERB is deployed to aggregate them into an enhanced edge $E^* \in \mathbb{R}^{4H\times 4W\times 3}$. Finally, we replace the blurred edge in $I_{\mathcal{P}}$ with the enhanced edges E^* to gain a reconstructed image $I_{\text{SR}} \in \mathbb{R}^{4H\times 4W\times 3}$ with refined and accurate edges.

For the discriminator, VGG-Discriminator [26] is chosen to determine whether the output image I_{SR} of the generator is real or fake, which is widely used in the SR domain.

B. Initial Feature Extraction

Due to the complex degradation in LR images, a significant amount of crucial high-frequency information is lost, making it difficult to achieve satisfactory results by directly enhancing the edges of LR images. Therefore, to provide exact information for subsequent modules to extract features, PSRM is built to process the LR image for generating Pre-SR images. The PSRM is an SR network that builds upon the strengths of SwinIR [41], which is a state-of-the-art solution in the SR domain, while also addressing its limitations. Although SwinIR can generate relatively clear images, its utilization of the L1 loss function may cause the generated SR images to appear overly smooth. This is due to the tendency of the network to generate images with pixel values closer to the average, resulting in the loss of high-frequency information and details. To address the limited ability of existing methods to accurately recover edges using only edge domain features, we extract features from both pixel and edge domains via the use of two convolutional layers. Let $F_c^1 = [f_{c1}^1, f_{c2}^1, f_{c3}^1]$ and $F_e^1 = [f_{e1}^1, f_{e2}^1, f_{e3}^1]$ denote the initial MS context features and edge features extracted by the corresponding convolutional layers. Taking edge features as an example, an edge image of size $4H \times 4W$ sequentially passes through three convolutional blocks, resulting in outputs of size $4H \times 4W$, $2H \times 2W$, and $H \times W$, respectively. Specifically, each convolutional block consists of three convolutional operators with kernel sizes of 1×1, 3×3, and 3×3. The only difference among the three blocks lies in the stride of the last 3×3 convolutional operator, which is 1, 2, and 2, respectively. The computation process for context features is the same as that for edge features. The convolutions in the pixel domain can extract rich context semantic features, which can guide the enhancement of edge features. Moreover, objects in remote sensing images usually cover a large span at the scale dimension, which makes it challenging for a single-scale feature to capture all the information needed for SR reconstruction. Therefore, the proposed convolutional layer can extract MS features, which enables it to adapt to various object sizes in remote sensing images.

For the edge extraction, there are several methods, such as Canny [47], HED [48], and EDTER [49]. Although the edge extraction algorithm can get accurate contours of objects, it cannot reflect the high-frequency information in noncontour regions, which is equally important for edge reconstruction. In contrast, many objects in remote sensing images do not have clear and well-defined target edges and we need the trend of object contour in the image. Therefore, the edge extraction method cannot meet the requirements of our proposed framework. The Laplacian is a second-order derivative-based edge-aware operator, which can extract the edges in an image while suppressing the smooth regions. Due to its nonlinear nature, it can enhance image edges and fine details without introducing artifacts or ringing effects. Therefore, the eight-directional Laplacian operator is chosen as the edge extraction that can represent the intensity of the change of each position in the image. The Laplacian operator $L(\cdot)$ of given image h can be expressed as

$$L(h) = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \otimes h \tag{1}$$

where $\otimes$ is the convolution operator.

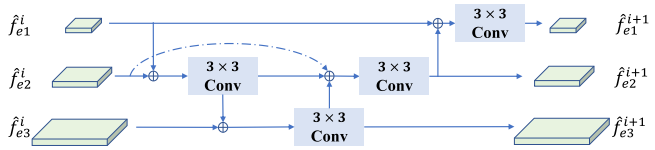

Fig. 2. Architecture of the BiEFFB and BiCFFB. $\oplus$ denotes the weighted sum operation. The LeakyReLU after each Conv layer is omitted for simplicity.

C. Edge Feature Enhanced Module

Due to the particular characteristics of remote sensing images, natural factors, such as clouds and fog often interfere with the objects and blur the edge features of the objects. In addition, since the complex background characteristics of remote sensing images, a large amount of noise is introduced in the feature extraction process. These issues lead to insufficient initial edge feature characterization ability. To address this problem, we design a bidirectional connection and weighted feature fusion block, named BiEFFB and BiCFFB. To preserve the spatial consistency of the two categories of features, we maintain the same architectural settings of convolution parameters. As shown in Fig. 2, BiEFFB or BiCFFB consists of four 3×3 convolutional layers with stride 1. Let $\hat{f}_{e1}^{i+1}$ denote the smallest size feature in the output of the ith EFEM, and it can be calculated as follows:

$$\hat{f}_{e1}^{i+1} = \text{LReLU}\left(\text{Conv}\left(\frac{w_1^{i+1}\cdot \hat{f}_{e1}^{i} + w_2^{i+1}\cdot \text{Resize}\left(\hat{f}_{e2}^{i+1}\right)}{w_1^{i+1} + w_2^{i+1} + \epsilon}\right)\right) \tag{2}$$

where $\text{LReLU}(\cdot)$ indicates the LeakyReLU activation function. $\text{Resize}(\cdot)$ represents the use of interpolation operations to make the dimensions consistent. w_1^{i+1} and w_2^{i+1} are two learnable weights that refer to the importance of each feature. The superscript $i+1$ indicates that the variable represents the weight for

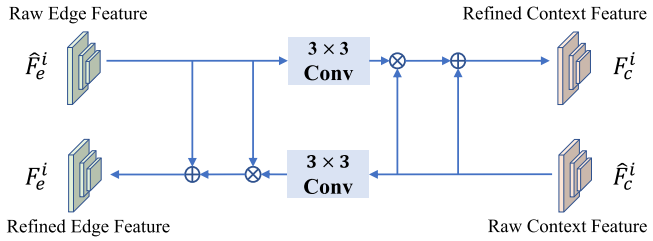

Fig. 3. Architecture of the CEIEB. Variables with the hat symbol indicate raw features without information exchange, and variables without the hat symbol indicate refined features after information exchange. $\otimes$ and $\oplus$ denote the element-wise multiplication and sum operation, respectively. The LeakyReLU after each Conv layer is omitted for simplicity.

the $i+1$th EFEM. ϵ is a small value to increase stability and is set to 10^{-6}. According to the same calculation method, we can obtain the enhanced features $\hat{F}_e^{i+1} = [\hat{f}_{e1}^{i+1}, \hat{f}_{e2}^{i+1}, \hat{f}_{e3}^{i+1}]$ and $\hat{F}_c^{i+1} = [\hat{f}_{c1}^{i+1}, \hat{f}_{c2}^{i+1}, \hat{f}_{c3}^{i+1}]$ from $\hat{F}_e^i$ and $\hat{F}_c^i$. The quality of the edge features is essential to determining the quality of the final edge enhancement images obtained by subsequent modules. To gradually convert the edge features containing noise into accurate features, N EFEMs are deployed as a cascaded module to fuse and exchange information for each context and edge feature. The analysis and discussion of the number of EFEM cascades are presented in Section IV-D4.

To overcome the limitations of current edge enhancement methods that only use edge features, we design a method that exploits the interplay between context semantic features present in the pixel domain and edge features to improve the accuracy of edge reconstruction. As shown in Fig. 3, the context semantic features from the context branch facilitate the edge branch's understanding of high-level context semantic information and contribute to the network's ability to learn edges better, resulting in more accurate edge reconstruction results. The information exchange process can be expressed as follows:

$$\begin{aligned} F_e^i &= \hat{F}_e^i + \text{LReLU}\left(\text{Conv}\left(\hat{F}_c^i\right)\right) \odot \hat{F}_e^i \\ F_c^i &= \hat{F}_c^i + \text{LReLU}\left(\text{Conv}\left(\hat{F}_e^i\right)\right) \odot \hat{F}_c^i \end{aligned} \tag{3}$$

where $\odot$ denotes the element-wise multiplication operation. In practical applications, we calculate each element contained in $\hat{F}_e^i$ and $\hat{F}_c^i$ according to (3). The results F_e^i and F_c^i can be used as input for the next EFEM. The use of multiple EFEMs allows for the progressive refinement of image features and the improvement of image quality over successive iterations. After N EFEMs enhancement, we can obtain the final results F_e^{N+1} and F_c^{N+1}, representing the enhanced edge and context feature, respectively.

D. Edge Restoration Block

Most SR methods only extract and utilize features at a single scale, which results in limited utilization of the image's information and consequently restricts the effectiveness of the model. To take advantage of MS features, we design the ERB, which can aggregate the enhanced MS edge features obtained from EFEMs to obtain an enhanced edge image that can reflect the reality edges of objects of different scales and classes. The structure of ERB is depicted in Fig. 4. Let $F_R = [f_{r1}, f_{r2}, f_{r3}]$ denote the output features of the residual connection with different scales in F_R. These features are resized to the same shape as the SR image and concatenated together to form a unified feature representation F_M. The process can be expressed as

$$F_R = F_e^{N+1} + F_e^1 = [f_{r1}, f_{r2}, f_{r3}] \tag{4}$$

$$F_M = \text{Concat}(f_{r1}, R_2(f_{r2}), R_4(f_{r3})) \tag{5}$$

where $\text{Concat}(\cdot)$ represents the concatenate operation in the channel dimension. R_2 and R_4 denote the $2\times$ upscale and $4\times$ upscale interpolation operation, respectively. After the scaling operation, all three features remain in the same feature dimension, where $f_{r1}, R_2(f_{r2}), R_4(f_{r3}) \in \mathbb{R}^{4H\times 4W\times C}$. C denotes the number of output channels. The output $F_M \in \mathbb{R}^{4H\times 4W\times 3C}$ will be used for further feature fusion.

Distinct from features present in the pixel domain, features extracted from the edge map contain limited information. Additionally, in the edge map, distinct objects may be represented by the same value, making it challenging to accurately distinguish between different objects during edge map reconstruction. To reduce the impact of noise and improve the accuracy of the reconstructed image, it is necessary to deeply explore the relationship between different regions of the input image. Coordinate attention (CA) [50] is employed to enhance the model's understanding of edge features, which integrates location information into the channel, enabling more precise localization and identification of the target of interest. By leveraging CA, our model gains a deeper understanding of the spatial context and improves its ability to accurately capture and utilize edge features for enhanced performance.

As shown in Fig. 4, the output of the CA is then passed through a convolution network consisting of three layers with kernel sizes of 1×1, 3×3, and 1×1, respectively. Each layer is followed by a LeakyReLU activation function. MS features are fused to produce the enhanced edge maps E^*. The process can be expressed as follows:

$$E^* = \text{FS}(\text{CA}(F_M)) \tag{6}$$

where $\text{CA}(\cdot)$ denotes the CA module and $\text{FS}(\cdot)$ represents the feature fusion module.

As shown in Fig. 1, to obtain the SR image I_{SR}, we add the desired amount of edge enhancement, denoted as $E^* - L(I_{\mathcal{P}})$, to the Pre-SR image $I_{\mathcal{P}}$. The plus and minus in Fig. 1 represent addition and subtraction operations, respectively, performed pixel by pixel when computing the SR image. This calculation process can be expressed as

$$I_{\text{SR}} = I_{\mathcal{P}} + E^* - L(I_{\mathcal{P}}). \tag{7}$$

E. Loss Functions in CEEGAN

In the SR field, L1 loss is a widely used loss function. It can reflect the difference between HR and SR images in the pixel domain. The L1 loss is calculated as follows:

$$\mathcal{L}_1 = \|I_{\text{HR}} - I_{\text{SR}}\|_1 \tag{8}$$

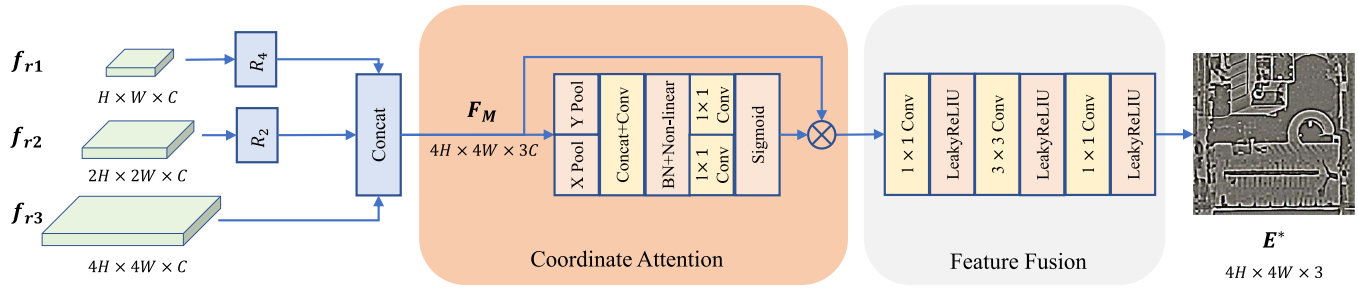

Fig. 4. Architecture of the ERB. R_2 and R_4 denote the $2\times$ upscale and $4\times$ upscale interpolation operation, respectively. $\otimes$ denotes the element-wise multiplication operation.

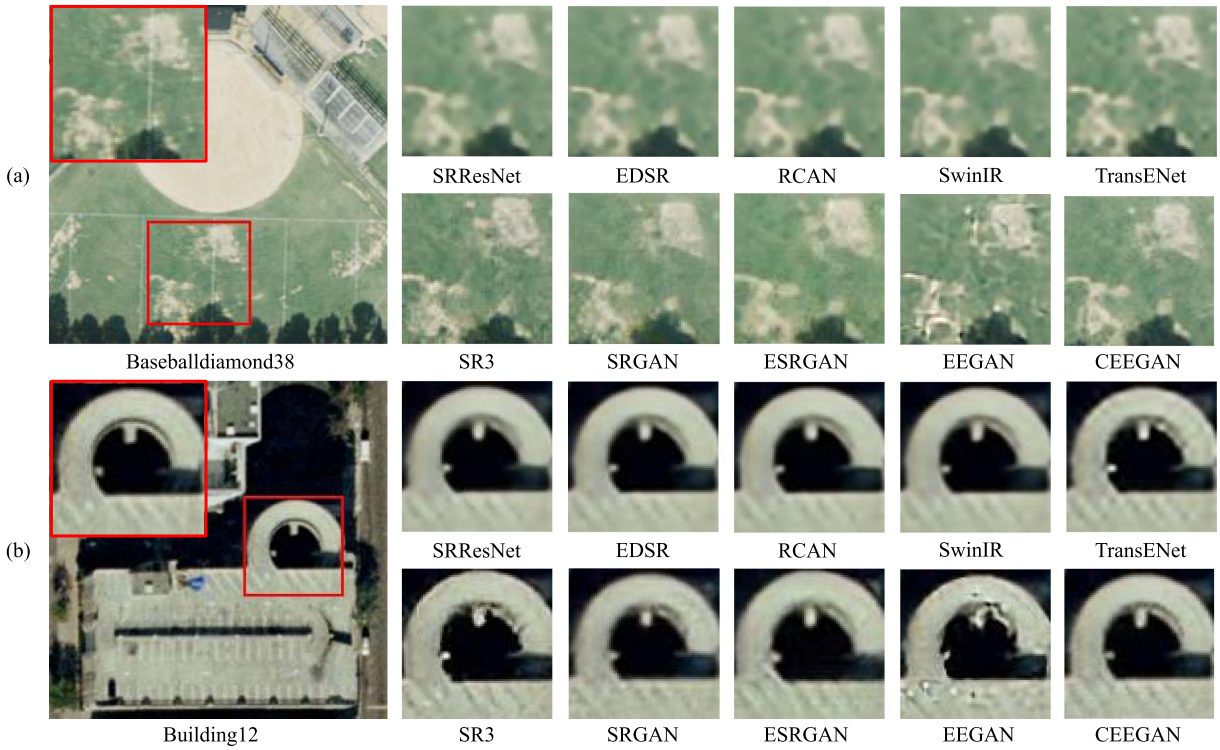

Fig. 5. Visualization comparison of different methods on UCMerced dataset. (a) Baseballdiamond38. (b) Building12. The GT images are zoomed in and displayed in the top-left corner.

where I_{HR} refer to the ground-truth image. The L1 loss is beneficial for optimizing PSNR, but the network will tend to output smooth results without sufficient high-frequency detail because the L1 loss and PSNR metrics are fundamentally inconsistent with the subjective evaluation of human observers. GAN-based methods typically use the sum of pixel loss and adversarial loss. In addition, to ensure the generation of high-frequency details, it is also necessary to utilize perceptual loss [51], which is usually a pretrained VGG network. Adversarial loss $\mathcal{L}_{\mathrm{adv}}$ and perceptual loss $\mathcal{L}_{\mathrm{per}}$ are defined as

$$\mathcal{L}_{\mathrm{adv}} = -\log\left(\mathcal{D}\left(I_{\mathrm{SR}}\right)\right) = -\log\left(\mathcal{D}\left(\mathcal{G}\left(I_{\mathrm{LR}}\right)\right)\right) \tag{9}$$

$$\mathcal{L}_{\mathrm{per}} = \sum_{l=1}^{n} \omega_l \left\|\phi_l(I_{\mathrm{HR}}) - \phi_l(I_{\mathrm{SR}})\right\|_2^2 \tag{10}$$

where $\phi_l(\cdot)$ represents the lth layer of the VGG19 network and ω_l refers to the weight of the lth layer. $\mathcal{G}(\cdot)$ and $\mathcal{D}(\cdot)$ refer to the generator and the discriminator, respectively.

Although the combination of the abovementioned loss functions performs better on natural images, it suffers from severe image degradation in the field of remote sensing images. This results in reconstructed images that still have noise and artifacts in the high-frequency detail parts. To solve this problem by strengthening the constraints in the high-frequency component of the image. We design the edge loss $\mathcal{L}_{\mathrm{Edge}}$, which is formulated as follows:

$$\mathcal{L}_{\mathrm{Edge}} = \left\|L(I_{\mathrm{HR}}) - L(I_{\mathrm{SR}})\right\|_1. \tag{11}$$

Finally, the total loss for the generator is given by

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_1 + \alpha\mathcal{L}_{\mathrm{adv}} + \beta\mathcal{L}_{\mathrm{per}} + \gamma\mathcal{L}_{\mathrm{Edge}} \tag{12}$$

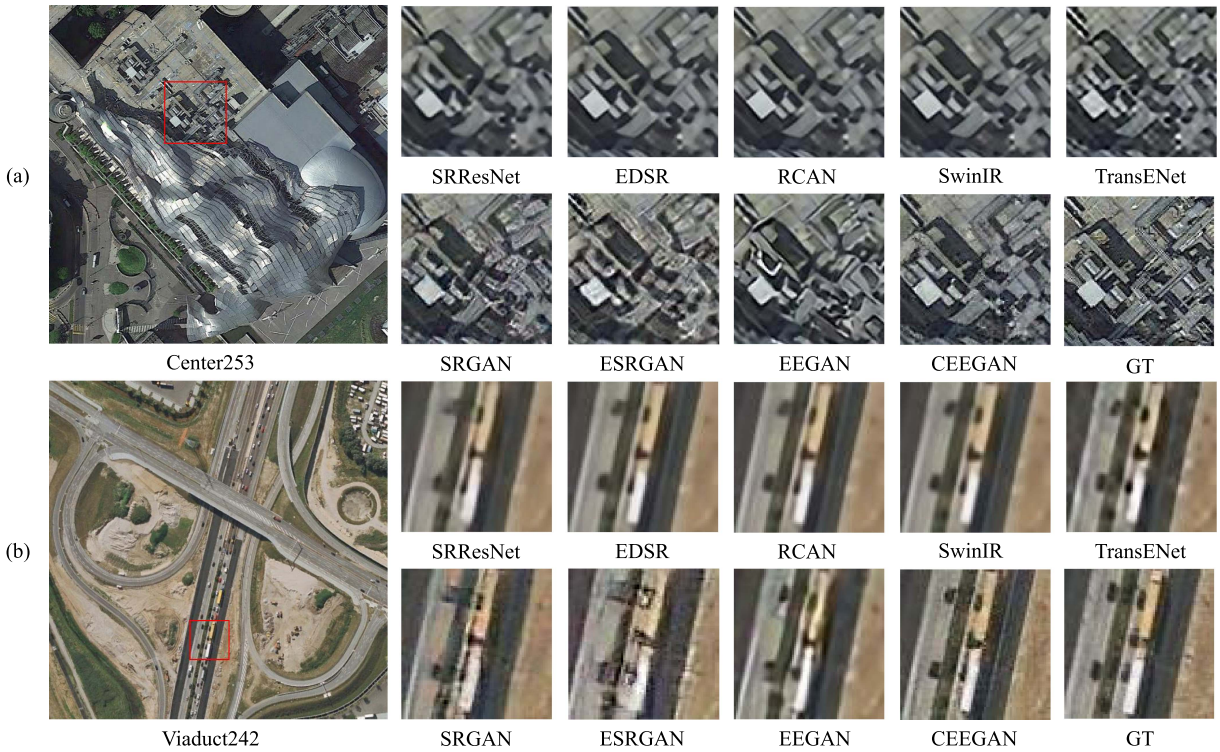

Fig. 6. Visualization comparison of different methods on AID dataset. (a) Center253. (b) Viaduct242.

where α, β, γ are the weighting parameters of each component, which is designed to balance the magnitudes of different losses. In our experiment, we set them to 0.1, 1, and 0.5.

For the discriminator, the loss widely used in other GAN-based methods is selected, which is calculated as

$$\mathcal{L}_{\mathcal{D}} = -\log\left(\mathcal{D}\left(I_{\mathrm{HR}}\right)\right) - \log\left(1 - \mathcal{D}\left(I_{\mathrm{SR}}\right)\right). \tag{13}$$

We train $\mathcal{D}$ by minimizing $\mathcal{L}_{\mathcal{D}}$, which allows it to distinguish whether the input image is synthesized by the $\mathcal{G}$.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: In this work, we select four public remote sensing datasets, including UCMerced [52], AID [53], NWPU-RESISC45 [54], and LoveDA [55] for SR, classification, and semantic segmentation. The UCMerced dataset contains 21 scene categories.[1] Each class has 100 images with a spatial resolution of one foot, and the size of all images is 256 × 256. We randomly select 80% of them as the training set and the rest as the validation set. The AID dataset contains 10 000 images in 30 classes of scenes.[2] All images are 600 × 600 in size. For the AID dataset, we randomly select 80% images as the training set and ten images per class as the validation set. The NWPU-RESISC45 remote sensing dataset is a large-scale public dataset for remote sensing image scene classification released by Northwestern Polytechnical University, which contains 45 different scenes. It contains a total of 31 500 images, each with a size of 256 × 256. The LoveDA dataset contains 5987 high SR images with 166 768 annotated objects from three different cities. It encompasses two domains, including urban and rural. Compared to existing datasets, the LoveDA dataset presents considerable challenges due to its MS objects and complex background samples.

2) Evaluation Metrics: To make a comprehensive comparison of the performance of the model, we select three commonly used evaluation metrics that focus on different aspects. The first metric is PSNR, which is an objective criterion for evaluating images. However, a higher PSNR does not necessarily correspond to better perceptual quality. To better simulate human visual perception, Zhang et al. [56] proposed learned perceptual image patch similarity (LPIPS), which measures the similarity of two images in a way that is more in line with human judgment. To validate the performance of models in real-world scenarios,

[1]All these 21 classes of UCMerced dataset: 1—Agricultural, 2—Airplane, 3—Baseballdiamond, 4—Beach, 5—Buildings, 6—Chaparral, 7—Denseresidential, 8—Forest, 9—Freeway, 10—Golfcourse, 11—Harbor, 12—Intersection, 13—Mediumresidential, 14—Mobilehomepark, 15—Overpass, 16—Parkinglot, 17—River, 18—Runway, 19—Sparseresidential, 20—Storagetanks, and 21—Tenniscourt.

[2]All these 30 classes of AID dataset: 1—Airport, 2—Bareland, 3—Baseballdiamond, 4—Beach, 5—Bridge, 6—Center, 7—Church, 8—Commercial, 9—Denseresidential, 10—Desert, 11—Farmland, 12—Forest, 13—Industrial, 14—Meadow, 15—Mediumresidential, 16—Mountain, 17—Park, 18—Parking, 19—Playground, 20—Pond, 21—Port, 22—Railwaystation, 23—Resort, 24—River, 25—School, 26—Sparseresidential, 27—Square, 28—Stadium, 29—Storagetanks, 30—Viaduct.

TABLE I
RESULTS OF COMPARISON WITH DIFFERENT METHODS ON THE UCMERCED DATASET

Class	PSNR-Oriented												GAN-Based							
	SRResNet[26]		EDSR[27]		RCAN[18]		SwinIR[41]		TransENet[42]		SR3[43]		SRGAN[26]		ESRGAN[28]		EEGAN[29]		CEEGAN	
	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS
1	28.39	0.3640	30.30	0.2804	30.83	0.2635	32.07	0.2349	29.42	0.2465	25.46	0.309	25.79	0.3109	24.70	0.3155	26.18	0.3529	29.27	**0.1990**
2	29.34	0.3195	29.79	0.2843	29.82	0.2866	30.46	0.2821	29.06	0.2874	26.33	0.2849	27.57	0.2760	25.50	0.2737	25.57	0.3833	28.56	**0.2657**
3	32.64	0.3318	32.88	0.3080	32.97	0.3100	33.17	0.3096	32.27	0.3043	29.37	0.2724	30.62	0.2673	29.73	0.2663	28.90	0.3610	31.20	**0.2502**
4	35.52	0.3158	35.63	0.2997	35.73	0.3017	35.99	0.2931	35.17	0.2852	29.82	0.2610	33.08	0.2703	32.62	0.2740	31.77	0.3207	33.50	**0.2626**
5	26.63	0.3100	27.00	0.2661	27.21	0.2643	27.98	0.2599	26.39	0.2715	23.81	0.2874	24.89	0.2828	24.14	0.2819	22.93	0.3759	26.42	**0.2411**
6	26.10	0.4386	26.30	0.3942	26.35	0.3961	26.69	0.3803	25.93	0.3984	23.31	0.2826	23.65	0.2954	23.04	0.2845	21.66	0.5216	24.26	**0.2581**
7	26.20	0.3451	26.70	0.2971	26.79	0.2971	27.53	0.2882	26.25	0.2907	23.78	0.2879	24.33	0.3063	23.44	0.3012	22.21	0.4046	25.86	**0.2474**
8	27.66	0.4578	27.90	0.4009	27.91	0.3987	28.13	0.3862	27.51	0.3973	24.17	0.3377	24.68	0.3652	24.41	0.3586	22.18	0.4883	25.35	**0.3283**
9	29.00	0.3263	30.05	0.2735	30.13	0.2737	30.71	0.2684	29.07	0.2770	25.98	0.2629	27.11	0.2729	25.92	0.2702	25.57	0.3744	28.57	**0.2280**
10	32.66	0.3555	32.88	0.3324	32.91	0.3344	33.11	0.3300	32.33	0.3233	28.93	0.2752	30.32	0.2755	29.58	0.2760	28.71	0.3634	30.82	**0.2545**
11	22.98	0.2881	23.70	0.2383	23.99	0.2351	24.75	0.2281	23.21	0.2465	21.00	0.2572	21.21	0.2657	20.60	0.2691	19.59	0.3588	23.12	**0.2163**
12	27.36	0.3324	27.80	0.2875	27.90	0.2875	28.51	0.2767	27.37	0.2833	24.91	0.2847	25.49	0.2974	24.53	0.3069	23.82	0.3935	26.89	**0.2542**
13	25.13	0.3980	25.41	0.3555	25.47	0.3552	26.00	0.3489	25.05	0.3436	22.38	0.3148	23.17	0.3327	22.33	0.3309	21.37	0.4359	24.33	**0.2839**
14	24.83	0.3658	25.46	0.3092	25.53	0.3108	26.15	0.3023	24.91	0.3092	22.39	0.3018	22.77	0.3165	21.88	0.3161	21.30	0.4164	24.76	**0.2575**
15	26.98	0.3169	28.04	0.2596	28.16	0.2590	28.91	0.2505	27.22	0.2787	24.15	0.2890	25.19	0.2849	23.78	0.2892	23.95	0.3915	27.02	**0.2337**
16	22.26	0.3510	23.11	0.2702	23.24	0.2673	24.04	0.2533	22.87	0.2603	20.33	0.2610	20.44	0.2864	19.49	0.2769	18.72	0.4178	22.35	**0.2332**
17	27.76	0.4176	27.98	0.3839	28.01	0.3816	28.16	0.3783	27.58	0.3790	24.50	0.3252	25.23	0.3342	24.88	0.3351	23.42	0.4303	25.86	**0.3075**
18	30.80	0.3233	32.02	0.2755	32.10	0.2755	32.72	0.2759	30.73	0.2847	27.28	0.2551	28.50	0.2602	26.93	0.2433	27.45	0.3589	29.91	**0.2421**
19	29.19	0.3828	29.41	0.3561	29.49	0.3587	29.76	0.3520	28.90	0.3462	25.85	0.3067	26.94	0.3195	26.24	0.3112	24.86	0.4179	27.57	**0.2833**
20	28.29	0.3278	28.58	0.2922	28.73	0.2927	29.26	0.2901	28.00	0.3002	24.90	0.3054	26.32	0.2900	25.24	0.2866	24.82	0.3802	27.45	**0.2610**
21	28.81	0.3432	29.46	0.2893	29.55	0.2885	30.08	0.2828	28.85	0.2928	26.19	0.2936	26.84	0.3035	25.81	0.3030	24.97	0.3931	28.28	**0.2484**
Avg	28.02	0.3529	28.59	0.3073	28.71	0.3066	29.25	0.2986	28.00	0.3051	24.99	0.2884	25.91	0.2959	24.99	0.2938	24.28	0.3972	27.21	**0.2550**

The best result of each metric is in bold font.

we selected the natural image quality evaluator (NIQE) as the no-reference metric. It should be noted that the lower LPIPS and NIQE values represent the better effect of SR reconstruction. For the classification experiment, top-1 accuracy is used as the metric. In the semantic segmentation experiment, we select three general evaluation indicators: overall accuracy (OA), mean intersection over union (mIoU), and F1-score.

B. Implementation Details and Experimental Environment

For training, the input patches are cropped in size of 128×128 pixels with a batch size of 16. Meanwhile, we use random rotation and horizontal flipping to augment the training samples. For optimization, we use Adam optimizer by setting $\beta_1 = 0.9, \beta_2 = 0.99$, and $\epsilon = 10^{-8}$. We train the PSRM in IFE for 1600 epochs and fine-tune the whole CEEGAN for 400 epochs. The learning rate is initialized 2×10^{-4} and decreases to half every 400 epochs and 100 epochs for two stages. In our experiments, the raw images in each dataset are treated as real HR references and corresponding LR images are generated by Gaussian blur and bicubic interpolation to construct HR–LR pairs for training and evaluation. We first use the degraded image as the input of different SR reconstruction methods for the classification and semantic segmentation experiments. Then, we evaluate the corresponding results by the same classification or semantic segmentation model. The proposed method is implemented by the PyTorch framework under Ubuntu18.04 and CUDA10.2. All experiments are run on NVIDIA 2080Ti GPUs.

C. Performance of Human Visual Perception and Machine Vision Applications

In this section, we compare the proposed methods with state-of-the-art SR methods, including PSNR-Oriented approaches: SRResNet [26], EDSR [27], RCAN [18], SwinIR [41], TransEnet [42], SR3 [43], GAN-based methods: EEGAN [29], SRGAN [26], ESRGAN [28]. EEGAN and TransENet are proposed for remote sensing images, and other methods are proposed for natural images. For a fair comparison, we train these models using the same training datasets.

1) Quantitative Results: Due to the different learning difficulties of different scenarios, which tend to lead to large fluctuations in indicators. To make a fair comparison, we calculate PSNR and LPIPS separately for each category and report the average scores. Tables I and II show the results of different methods for upscale $\times 4$ on the UCMerced and AID dataset, respectively. From a macro perspective, PSNR-oriented methods dominate PSNR and GAN-based methods outperform by a significant margin in terms of LPIPS. Due to the addition of high-frequency information by GAN to generated images aims to produce more realistic images, it may result in a decrease in the PSNR. CEEGAN can outperform other methods compared with in terms of LPIPS in both two datasets, which means it can get better visual quality results. This is because EFEM and edge loss, designed for edge enhancement in CEEGAN, make the model pay more attention to edge details when restoring images. Meanwhile, ERB can fuse MS features so that refined details can be obtained on both large and small objects. The

TABLE II
RESULTS OF COMPARISON WITH DIFFERENT METHODS ON THE AID DATASET

Class	PSNR-Oriented										GAN-Based							
	SRResNet[26]		EDSR[27]		RCAN[18]		SwinIR[41]		TransENet[42]		SRGAN[26]		ESRGAN[28]		EEGAN[29]		CEEGAN	
	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS
1	28.58	0.3880	28.87	0.3665	28.89	0.3661	28.94	0.3682	28.03	0.3704	26.76	0.3617	24.95	0.3652	26.36	0.3948	26.72	**0.3349**
2	38.96	0.3274	39.13	0.3239	39.15	0.3213	39.14	0.3234	37.70	0.3243	36.84	0.3266	34.43	0.3438	37.48	0.3212	36.16	**0.3175**
3	27.80	0.3528	28.03	0.3375	28.07	0.3361	28.06	0.3374	29.60	0.3409	25.58	0.3625	24.44	0.3536	27.90	0.3598	26.03	**0.3188**
4	27.86	0.3887	27.97	0.3779	27.98	0.3751	27.98	0.3767	32.13	0.3756	25.36	0.3950	24.81	0.3822	31.36	0.3845	26.48	**0.3221**
5	34.26	0.3631	34.55	0.3567	34.61	0.3580	34.63	0.3593	33.45	0.3583	30.90	0.4376	30.05	0.3667	32.00	0.3719	32.30	**0.3195**
6	24.18	0.3537	24.45	0.3348	24.50	0.3339	24.52	0.3356	27.36	0.3344	22.80	0.3238	21.39	0.3174	26.02	0.3609	23.05	**0.2930**
7	22.67	0.3693	22.96	0.3448	22.96	0.3445	22.99	0.3448	24.57	0.3447	21.08	0.3484	19.96	0.3334	22.62	0.3848	21.44	**0.3152**
8	23.30	0.3806	23.47	0.3628	23.48	0.3592	23.49	0.3639	26.41	0.3600	21.74	0.3557	20.54	0.3602	24.31	0.4000	21.95	**0.3325**
9	23.33	0.4287	23.61	0.3998	23.63	0.3999	23.63	0.4019	23.05	0.3938	20.86	0.3727	19.71	0.3734	21.00	0.4407	21.51	**0.3348**
10	36.57	0.4169	36.65	0.4125	36.65	0.4114	36.67	0.4129	35.69	0.4090	35.04	0.3734	32.89	0.3817	35.38	0.4120	34.30	**0.3704**
11	30.53	0.3537	30.75	0.3383	31.12	0.3261	31.14	0.3235	33.14	0.3392	28.79	0.3619	27.15	0.3644	32.49	0.3588	28.78	**0.3090**
12	29.76	0.4351	29.83	0.4113	29.87	0.4022	29.87	0.4114	29.07	0.4307	26.12	0.3891	24.45	0.4010	28.21	0.4482	27.12	**0.3592**
13	22.52	0.3838	22.71	0.3555	22.73	0.3543	22.74	0.3576	26.95	0.3574	20.96	0.3580	20.00	0.3598	25.11	0.4012	21.12	**0.3272**
14	33.81	0.4034	33.99	0.3910	34.03	0.3881	34.02	0.3904	32.98	0.3986	29.79	0.4263	28.13	0.4382	32.82	0.4077	31.09	**0.3622**
15	25.51	0.4324	25.74	0.4107	25.78	0.4068	25.78	0.4096	24.99	0.4149	23.07	0.3686	21.79	0.3547	23.38	0.4417	23.68	**0.3363**
16	25.42	0.3827	25.49	0.3728	25.50	**0.3715**	25.49	0.3724	29.01	0.3840	23.70	0.3922	22.30	0.3912	27.74	0.4005	23.54	0.3826
17	29.32	0.3706	29.58	0.3555	29.60	0.3529	29.61	0.3548	28.75	0.3589	27.07	0.3601	25.73	0.3584	26.90	0.3897	27.33	**0.3345**
18	25.34	0.3046	26.24	0.2745	26.43	0.2690	26.39	0.2721	25.40	0.2725	22.63	0.3102	21.32	0.3026	22.59	0.3290	24.34	**0.2575**
19	32.00	0.3359	32.75	0.3084	32.80	0.3059	32.77	0.3089	31.47	0.3140	29.71	0.3176	27.72	0.326	29.93	0.3448	30.00	**0.2699**
20	30.03	0.3714	30.21	0.3600	30.24	0.3580	30.24	0.3598	29.51	0.3633	27.31	0.4204	26.03	0.3857	28.20	0.3840	27.92	**0.3189**
21	22.79	0.3747	22.99	0.3598	23.03	0.3596	23.00	0.3616	26.28	0.3658	21.15	0.4169	20.41	0.3828	24.40	0.3863	21.47	**0.3269**
22	28.92	0.3734	29.38	0.3454	29.46	0.3436	29.48	0.3438	28.30	0.3434	26.82	0.3446	24.63	0.3608	26.44	0.3873	27.27	**0.3074**
23	28.89	0.3731	29.20	0.3579	29.21	0.3578	29.23	0.3579	28.32	0.3638	26.70	0.3533	25.25	0.3658	26.80	0.3853	27.04	**0.3218**
24	31.97	0.3962	32.09	0.3867	32.11	0.3850	32.12	0.3846	31.26	0.3899	29.37	0.3764	27.68	0.3814	30.40	0.4058	29.52	**0.3398**
25	20.10	0.4073	20.25	0.3820	20.26	0.3813	20.27	0.3841	25.16	0.3863	18.62	0.3708	17.86	0.3821	23.15	0.4198	18.89	**0.3348**
26	24.28	0.4896	24.35	0.4713	24.38	0.4667	24.37	0.4686	25.88	0.4924	22.11	0.3895	21.11	0.4081	25.06	0.4959	22.40	**0.3694**
27	27.11	0.3429	27.36	0.3263	27.40	0.3247	27.41	0.3260	29.47	0.3338	25.26	0.3099	23.68	0.3051	27.79	0.3563	25.59	**0.2847**
28	29.27	0.3311	29.87	0.3063	29.90	0.3075	29.96	0.3085	28.95	0.3111	27.56	0.3171	25.69	0.3142	27.17	0.3384	28.14	**0.2794**
29	27.62	0.3888	27.89	0.3725	27.90	0.3731	27.89	0.3737	27.08	0.3742	25.73	0.3618	24.11	0.3609	25.60	0.3932	26.03	**0.3219**
30	27.90	0.3571	28.41	0.3339	28.45	0.3315	28.43	0.3354	27.37	0.3461	25.47	0.3247	23.60	0.3254	25.59	0.3700	25.78	**0.2950**
Avg	28.02	0.3792	28.29	0.3612	28.34	0.3590	28.34	0.3610	28.91	0.3651	25.83	0.3642	24.39	0.3615	27.47	0.3892	26.23	**0.3232**

The best result of each metric is in bold font.

clearer the edge details are, the more realistic image can be. Although EEGAN also focuses on improving image quality through edge enhancement, our method further incorporates edge domain constraints and context features to guide edge reconstruction, resulting in better performance in terms of LPIPS. It can be observed that CEEGAN can achieve the highest PSNR compared with other GAN-based methods on the UCMerced dataset.

2) Qualitative Comparison: To directly compare the reconstruction results obtained by different methods, Baseballdiamond38 and Building12 from the UCMerced dataset and Center253 and Viaduct24 from the AID dataset are selected to show the details of the SR images. As shown in Fig. 5 and Fig. 6, these SR images show that the CEEGAN can obtain more realistic reconstruction results at the edge of objects, such as roads and buildings. To intuitively demonstrate the effectiveness of our method on edge enhancement, we compare the images before and after edge enhancement, their response edge maps, and the difference between the two edges, as shown in Fig. 7. From Fig. 7(j), it can be seen the differences between E^* and $L(I_{\mathcal{P}})$ are mainly focused on the object edges, which proves that the region of interest of CEEGAN is the edge. The addition of high-frequency information can make the image more realistic and comfortable for human vision.

3) Classification Results on UCMerced and NWPU-RESISC45 Datasets: To assess the application performance of CEEGAN in real scenarios, we conduct image classification experiments over the UCMerced and NWPU-RESISC45 datasets. First, we trained a ResNet-34 [44] image classification network over the original images in both datasets. Subsequently, we applied bicubic interpolation and Gaussian blur downsampling on the HR images with a size of 256 $\times$ 256 to generate LR images with a size of 64 $\times$ 64. We then utilize various SR methods to obtain the corresponding reconstruction images. Finally, we input the reconstructed images into the image classification network to obtain the classification result. As shown in Table III, the classification task performs the best over the SR images obtained by CEEGAN. Compared to SwinIR and SR3, which ranked second in the evaluation, CEEGAN can achieve significant improvements of 0.58% and 1.91% on the UCMerced

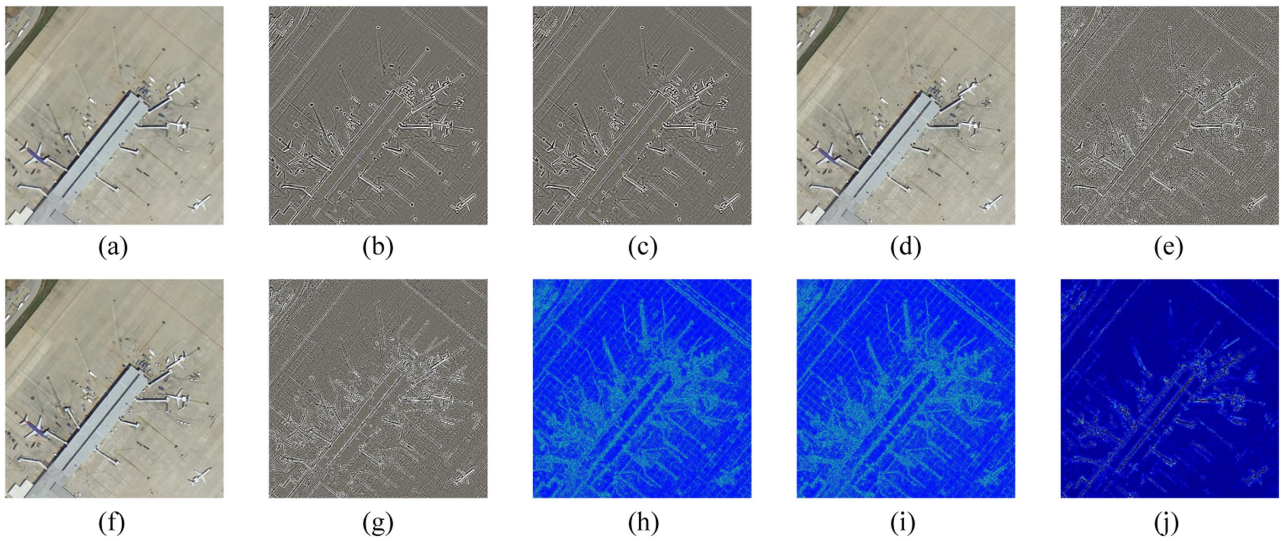

Fig. 7. Comparison of images before and after edge enhancement is implemented. (a) Pre-SR image $I_{\mathcal{P}}$. (b) Edge map of Pre-SR. $L(I_{\mathcal{P}})$ (c) Enhanced edge map E^*. (d) SR image I_{SR}. (e) Edge map of SR $L(I_{\text{SR}})$. (f) GT image I_{HR}. (g) Edge map of GT $L(I_{\text{GT}})$. (h) Difference of (b) and (g) $|L(I_{\text{HR}}) - L(I_{\mathcal{P}})|$. (i) Difference of (c) and (g) $|E^* - L(I_{\mathcal{P}})|$. (j) Difference of (b) and (c) $|E^* - L(I_{\mathcal{P}})|$.

TABLE III
CLASSIFACTION RESULTS ON UCMERCED AND NWPU-RESISC45 DATASET

	Bicubic	SRResNet	EDSR	RCAN	SwinIR	TransENet	SR3	SRGAN	ESRGAN	EEGAN	CEEGAN	HR
UCMerced	92.67	96.28	96.51	96.40	97.09	96.74	96.63	96.63	95.93	96.63	**97.67**	97.94
NWPU	88.13	94.10	95.33	95.11	94.95	94.44	96.35	96.05	95.43	95.78	**96.86**	99.52

The best result of each metric is in bold font.

TABLE IV
SEMANTIC SEGMENTATION RESULTS ON LOVEDA DATASET

	Bicubic	SRResNet	EDSR	RCAN	SwinIR	TransENet	SRGAN	ESRGAN	EEGAN	CEEGAN	HR
OA	59.91	58.20	60.48	60.41	61.71	61.84	61.89	59.48	53.56	**62.85**	68.98
mIoU	37.73	39.26	40.69	40.87	41.79	39.49	41.10	38.99	30.75	**41.93**	51.24
F1-score	53.50	55.48	57.00	57.25	58.12	55.38	57.34	55.40	45.18	**58.29**	66.95

The best result of each metric is in bold font.

and NWPU-RESISC45 datasets, respectively. Furthermore, it is worth noting that our method is capable of generating results that closely resemble HR images, indicating that our method can effectively preserve essential image features and produce highly realistic images. These advantages highlight the effectiveness and potential of our proposed method in practical applications.

4) Semantic Segmentation Results on LoveDA Dataset: To further verify the practicality of our method, we have designed a comparative experiment for the semantic segmentation task. The specific experimental process is similar to the image classification task. The downsampled images are reconstructed by different SR models. Then, they are input into the deeplabv3 [57] model to compare the segmentation results. It should be noted that the semantic segmentation models are trained on the LoveDA dataset, and the SR models are trained on the UCMerced dataset. Table IV lists the semantic segmentation evaluation results of SR images obtained by different SR models on the LoveDA dataset. Fig. 8 visualizes the semantic segmentation results. Our method can achieve the best performance not only in the classification task but also in semantic segmentation, with the highest values in OA, F1-score, and mIoU. This can be attributed to the ability of edge loss to constrain the edge graph, allowing the generation of accurate high-frequency information, which enable CEEGAN to achieve results far beyond EEGAN on semantic segmentation experiments. The result proves that CEEGAN can not only improve the model's understanding of the entire image but also improve the model's ability to understand images at the pixel scale.

5) Comparative Results of Real-World Scenarios SR: In real-world scenarios, the degradation process of images may not align with the assumptions made in creating SR datasets. Therefore, we select the no-reference metric NIQE to evaluate the performance of different models on a real-world scenario image from the WorldView-3 satellite. Fig. 9 shows the SR images reconstructed by different SR models along with their corresponding NIQE values. It can be observed that CEEGAN can achieve the lowest NIQE value, indicating that CEEGAN

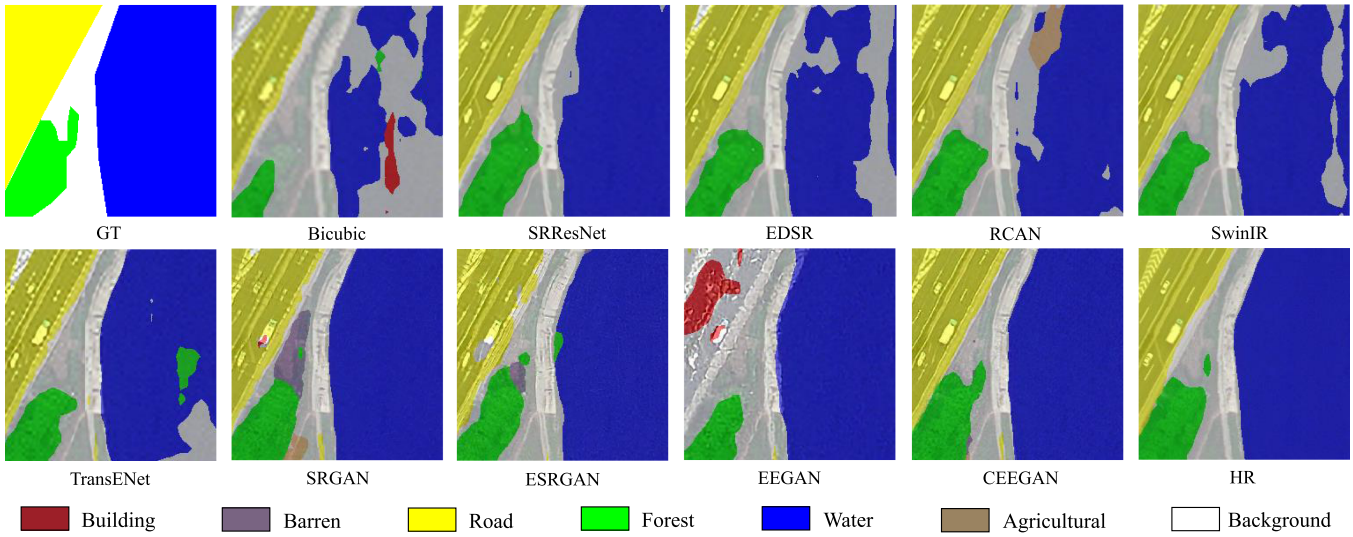

Fig. 8. Visualization comparison of reconstruction results using different SR methods in semantic segmentation experiments.

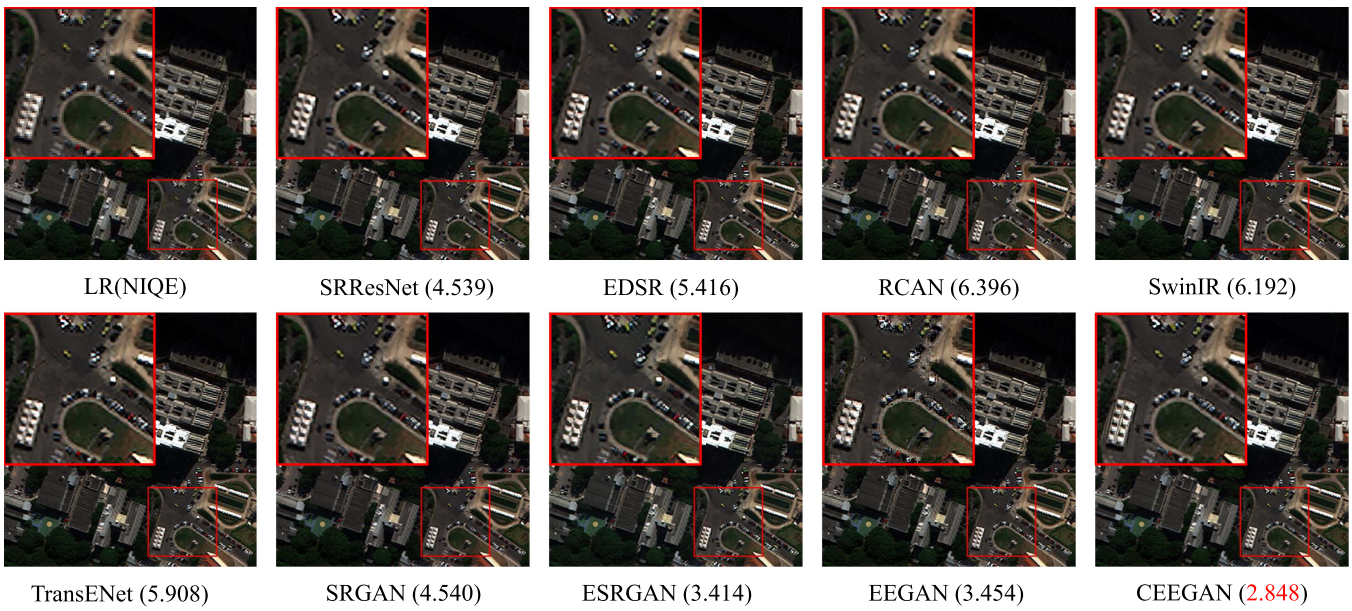

Fig. 9. Comparison of SR reconstruction results on real-world scene images using different methods. Red indicates the model that achieved the lowest NIQE.

can produce more natural results when faced with unknown degraded images.

6) Model Complexity Comparison: To compare the running efficiency of different models, we have compared the parameters, floating point operations (FLOPs), runtimes, LPIPS, and classification accuracy on the UCMerced dataset. Although our primary objective is to optimize the performance of the SR model for machine vision applications in remote sensing images, the results presented in Table V indicate that our model can also offer advantages in terms of parameter numbers, FLOPs, and run time. In addition, it outperforms some recent models for remote sensing image SR, such as TransENet and EEGAN, in terms of runtime. Moreover, compared to the state-of-the-art diffusion model in the field of SR, SR3, CEEGAN demonstrates advantages in terms of parameters, FLOPs, and runtimes. Although CEEGAN may exhibit a gap in terms of model efficiency compared to earlier algorithms like ESRGAN, it can achieve superior results in terms of LPIPS and practical applications by leveraging larger models and more advanced architectural designs. The comparative experimental results on model complexity confirm that CEEGAN acquires a better balance between complexity and accuracy and has more potential for practical applications in real remote sensing scenarios.

D. Ablation Studies

In this section, we design a series of experiments on the UCMereced dataset to validate the effectiveness and necessity of

TABLE V
QUANTITATIVE COMPARISON OF PARAMETERS, FLOPS, RUNTIMES, LPIPS, AND CLASSIFICATION ACCURACY FOR DIFFERENT METHODS

	SRResnet	EDSR	RCAN	SwinIR	TransENet	SRGAN	ESRGAN	EEGAN	SR3	CEEGAN
Parameters(M)	1.52	43.09	15.59	11.90	37.46	1.52	16.70	7.38	97.81	14.67
FLOPs(G)	10.39	205.83	65.25	50.55	48.23	10.39	73.43	78.48	179.33	149.67
Run Times(s)	0.0058	0.0515	0.1242	0.1392	0.1951	0.0058	0.0565	0.5530	90.89	0.1480
LPIPS	0.3529	0.3073	0.3066	0.2986	0.3051	0.2959	0.2938	0.3972	0.2884	0.2550
Accuracy	92.67	96.28	96.51	96.40	97.09	96.74	95.43	96.63	96.63	97.67

TABLE VI
QUANTITATIVE COMPARISON OF DIFFERENT COMPONENTS IN EFEM

Exp case	BiEFFB	BiCFFB	CEIEB	PSNR	LPIPS
I	×	×	×	27.88	0.2681
II	✓	×	×	26.83	0.2658
III	✓	✓	×	26.89	0.2638
IV	✓	✓	✓	27.21	0.2550

TABLE VII
QUANTITATIVE COMPARISON OF DIFFERENT COMPONENTS IN ERB

Exp case	MS	CA	PSNR	LPIPS
I	×	×	26.44	0.2857
II	✓	×	26.14	0.2832
III	×	✓	27.00	0.2660
IV	✓	✓	27.21	0.2550

each component in our proposed method. All models are trained with the same settings.

1) Effect of Each Part in EFEM: We first investigate the impact of each component in EFEM, including the BiEFFB, BiCFFB, and CEIEB. Table VI lists the evaluation results of different settings. The result of case IV is 4.03% higher than the result of case II on LPIPS, which indicates the guiding role of the semantic branch on the edge branch. Case III adds semantic branch to case II. Instead of exchanging information between any pair of BiEFFB and BiCFFB, it takes place once in the ERB. The results can achieve 0.75% improvement compared with case II on LPIPS, which can demonstrate the effectiveness of our proposed edge enhancement module based on context guidance. By comparing the results of cases III and IV, it can be concluded that the addition of CEIEB improves LPIPS by 3.30%, which indicates the necessity for semantic communication after each edge enhancement.

2) Effect of Each Part in ERB: We conduct the ablation study on the MS architectural and CA mechanism in ERB, as shown in Table VII. Compared to case IV, case I removes the MS structure and CA mechanism in ERB, and only utilizes the edge features from the maximum scale as input for edge reconstruction, which leads to a decrease of 0.0307 in LPIPS. Case II adds the MS architectural into case I. However, due to the absence of CA, which helps the model understand the importance of different channels and coordinates, the reconstruction results only improved by 0.0025. Case III adds the CA mechanism into case I, while still using only the maximum size features. Although it resulted in a 0.0197 improvement in LPIPS, using single-scale features cannot fully represent the objects that exist in the image.

TABLE VIII
QUANTITATIVE COMPARISON OF DIFFERENT LOSS FUNCTIONS

Exp case	Edge Enhancement	L1-Loss	GAN-Loss	Edge-Loss	PSNR	LPIPS
I	✓	✓	×	×	28.48	0.2721
II		✓	✓	×	26.95	0.2663
III		✓	×	✓	29.24	0.3064
IV		✓	✓	✓	27.21	0.2550
V	×	✓	×	×	29.25	0.3087
VI		✓	✓	×	27.30	0.2699
VII		✓	×	✓	29.08	0.3054
VIII		✓	✓	✓	27.88	0.2681

Compared to the combined usage of MS and CA in case IV, there is still a 0.0110 gap in maximizing the utilization of effective image information for edge reconstruction.

3) Effect of GAN-Loss and Edge Loss: To verify the effect of edge loss with and without EFEM. As listed in Table VIII, the cases I–IV are different loss function combinations with EFEM and ERB, and cases V–VIII remove the whole process of edge enhancement. The GAN-Loss includes adversarial loss and perceptual loss as presented in (9) and (10). The data from cases II and VI show that the addition of GAN-Loss causes the decrease of both PSNR and LPIPS, which is consistent with the purpose of adding GAN-Loss. The data from cases III and VII show that the addition of edge loss also causes a decrease in PSNR and LPIPS without EFEM, while its effect is more limited compared to that of GAN-Loss. When EFEM is available, the addition of edge loss can improve both PSNR and LPIPS. This result indicates that edge loss can work together with EFEM to recover more realistic image edge details in the absence of GAN-Loss. However, since edge loss is still essentially an L1 loss, the network is not optimized in the direction of conforming to human visual habits. Cases IV and VIII demonstrate that the combined use of both edge loss and GAN-Loss can make the model optimal at LPIPS with or without EFEM. This also implies that the network can conform itself to human visual habits and be suitable for practical machine vision applications by learning how to incorporate high-frequency information that is difficult to recover from L1 loss during the reconstruction process under the constraints of these two loss functions.

4) Number of EFEM: To examine how the number of EFEMs affects the effect of edge reconstruction and the performance of the model, ablation experiments are carried out and the results are shown in Table IX. Our experimental results show that increasing the number of EFEMs can improve the model performance. As shown in Fig. 10, the benefits of adding the

TABLE IX
QUANTITATIVE COMPARISON OF DIFFERENT NUMBERS OF EFEMS

Number of EFEMs	PSNR	LPIPS
0	27.88	0.2681
1	27.31	0.2593
3	27.22	0.2550
5	27.21	0.2532
7	27.10	0.2524

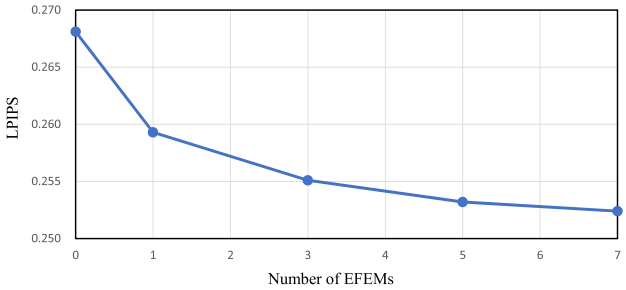

Fig. 10. Performance of LPIPS with the different numbers of EFEMs.

same number of EFEMs gradually decrease, while the increase in parameter count is fixed. Although EFEM can improve the model performance, its excessive use may increase model complexity and result in a potential loss of effectiveness, leading to overfitting and lengthier training times. Therefore, we decided to limit the number of EFEMs used in practice to three, as it strikes a balance between model performance and complexity.

V. CONCLUSION

In this work, a context aware EEGAN is proposed for remote sensing SR. To address the limitations of existing edge reconstruction methods, we propose a new edge enhancement method that simultaneously exploits contextual semantic features with edge features and maintains information exchange during the gradual enhancement process. To maintain the accuracy of the high-frequency information, the edge loss function is designed to constrain the generated edges in the edge domain.

We compare the SR results on the UCMerced and AID datasets with the current advanced methods and achieve the best results on the LPIPS. The ablation experiments prove the effectiveness of each part in CEEGAN and the Edge loss function. To further demonstrate the suitability of our results for practical remote sensing applications, classification, and segmentation experiments are conducted on the UCMerced and LoveDA datasets, respectively. CEEGAN achieves the best results on multiple evaluation metrics, which proves that our method is more suitable for the practical application of remote sensing image SR.

REFERENCES

[1] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[2] R. Dian, A. Guo, and S. Li, "Zero-shot hyperspectral sharpening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12650–12666, Oct. 2023.

[3] R. Dian, T. Shan, W. He, and H. Liu, "Spectral super-resolution via model-guided cross-fusion network," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3238506.

[4] L. He, W. Zhang, J. Shi, and F. Li, "Cross-domain association mining based generative adversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7770–7783, 2022.

[5] X. Guan, F. Li, X. Zhang, M. Ma, and S. Mei, "Assessing full-resolution pansharpening quality: A comparative study of methods and measurements," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6860–6875, Jul. 2023.

[6] S. Mei et al., "Lightweight multiresolution feature fusion network for spectral super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5501414.

[7] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2021, Art. no. 5509612.

[8] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5539014.

[9] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[10] J. Wang, F. Li, and H. Bi, "Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 4707013.

[11] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[12] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[13] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-Based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 4403718.

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

[15] R. Schultz and R. Stevenson, "A Bayesian approach to image expansion for improved definition," *IEEE Trans. Image Process.*, vol. 3, no. 3, pp. 233–242, May 1994.

[16] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 561–568.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5401410.

[24] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023.

[25] 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. 2022, Art. no. 5407915.

[26] 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.
[27] 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.
[28] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 63–79.
[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] 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. 2022, Art. no. 5624715.
[31] R. Dong, L. Zhang, and H. Fu, "RRSGAN: Reference-based super-resolution for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5601117.
[32] M. S. Moustafa and S. A. Sayed, "Satellite imagery super-resolution using squeeze-and-excitation-based GAN," *Int. J. Aeronautical Space Sci.*, vol. 22, no. 6, pp. 1481–1492, Dec. 2021.
[33] 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, Nov. 2018, Art. no. 1700.
[34] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5610819.
[35] P. Yi, Z. Wang, K. Jiang, J. Jiang, T. Lu, and J. Ma, "A progressive fusion generative adversarial network for realistic and consistent video super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2264–2280, May 2022.
[36] Z. Wang, K. Jiang, P. Yi, Z. Han, and Z. He, "Ultra-dense GAN for satellite imagery super-resolution," *Neurocomputing*, vol. 398, pp. 328–337, Jul. 2020.
[37] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014.
[38] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021.
[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[40] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
[41] 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. Workshops*, 2021, pp. 1833–1844.
[42] 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. 2022, Art. no. 5615611.
[43] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2022.
[44] 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.
[45] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.
[46] 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.
[47] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[48] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.
[49] M. Pu, Y. Huang, Y. Liu, Q. Guan, and H. Ling, "EDTER: Edge detection with transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1392–1402.
[50] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.
[51] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[52] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.
[53] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[54] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[55] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, 2021.
[56] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[57] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

Zhihan Ren received the B.S. degree in information engineering from the College of Communication Engineering, Jilin University, Changchun, China, in 2023. He is currently working toward the M.S. degree in information and communication engineering with the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China.

His research interests include deep learning and remote sensing image understanding and processing.

Lijun He received the B.S. degree in information engineering and Ph.D. degree in information and communications engineering from the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008 and 2016, respectively.

She is currently an Associate Professor with the School of Information and Communications Engineering, Xi'an Jiaotong University. Her research interests include video communication and transmission, video analysis, processing, and compression techniques.

Jichuan Lu received the master's degree in communication and information system from the School of Communication and Information Engineering, Xidian University, Xi'an, China, in 2011.

He is currently a Senior Expert with China Mobile Communications Group, Xi'an, China, mainly engaged in research in the fields of cloud computing and AI.