EDiffSR: An Efficient Diffusion Probabilistic Model for Remote Sensing Image Super-Resolution

Yi Xiao, Qiangqiang Yuan, *Member, IEEE,* Kui Jiang, *Member, IEEE,* Jiang He, *Graduate Student Member, IEEE,* Xianyu Jin, and Liangpei Zhang, *Fellow, IEEE.*

***Abstract*—Recently, convolutional networks have achieved remarkable development in remote sensing image Super-Resoltuion (SR) by minimizing the regression objectives, *e.g.,* MSE loss. However, despite achieving impressive performance, these methods often suffer from poor visual quality with over-smooth issues. Generative adversarial networks have the potential to infer intricate details, but they are easy to collapse, resulting in undesirable artifacts. To mitigate these issues, in this paper, we first introduce Diffusion Probabilistic Model (DPM) for efficient remote sensing image SR, dubbed EDiffSR. EDiffSR is easy to train and maintains the merits of DPM in generating perceptual-pleasant images. Specifically, different from previous works using heavy UNet for noise prediction, we develop an Efficient Activation Network (EANet) to achieve favorable noise prediction performance by simplified channel attention and simple gate operation, which dramatically reduces the computational budget. Moreover, to introduce more valuable prior knowledge into the proposed EDiffSR, a practical Conditional Prior Enhancement Module (CPEM) is developed to help extract an enriched condition. Unlike most DPM-based SR models that directly generate conditions by amplifying LR images, the proposed CPEM helps to retain more informative cues for accurate SR. Extensive experiments on four remote sensing datasets demonstrate that EDiffSR can restore visual-pleasant images on simulated and real-world remote sensing images, both quantitatively and qualitatively. The code of EDiffSR will be available at https://github.com/XY-boy/EDiffSR**

***Index Terms*—Image super-resolution, diffusion probabilistic model, prior enhancement, remote sensing.**

I. INTRODUCTION

SUPER-Resolution (SR) is a long-standing issue and remains an active research topic in the area of remote sensing [10]. SR aims to reconstruct a high-resolution (HR) image with rich texture details from a low-resolution (LR) image [11] [12] [13]. Currently, SR has been widely explored in remote sensing applications, including land-cover mapping [14] [15], hyperspectral image fusion [16] [17], product reconstruction [?], [18], and vehicle tracking [19]. Due to the inherently ill-posed nature [20], [21], SR is challenging because the HR counterpart may be infinite in the solution space given an LR input. In particular, for large-scale earth observation scenarios, SR becomes more complicated owing to various degradations, such as atmospheric scattering and platform tremors. Therefore, developing an effective SR method to reconstruct high-quality images is indispensable and of practical importance.

This work was supported in part by the National Natural Science Foundation of China under Grant 42230108 and 61971319. (*Corresponding author: Qiangqiang Yuan*.)

Yi Xiao, Qiangqiang Yuan, Jiang He, and Xianyu Jin are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430000, China (e-mail: xiao_yi@whu.edu.cn; yqiang86@gmail.com; jiang_he@whu.edu.cn; jin_xy@whu.edu.cn).

Kui Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150000, China (e-mail: kuijiang_1994@163.com).

Liangpei Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430000, China (e-mail: zlp62@whu.edu.cn).

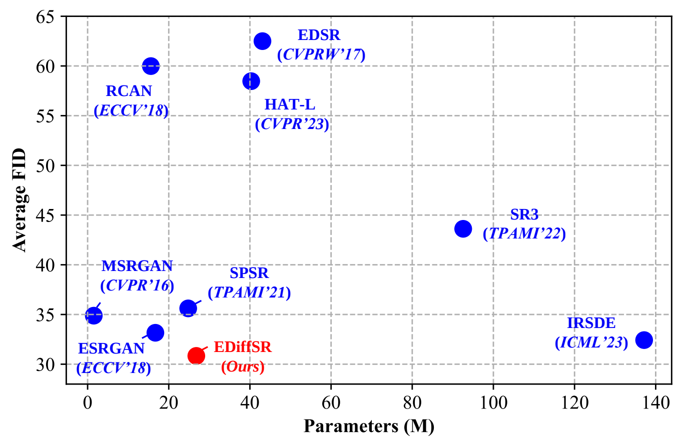

Fig. 1. The relationship between FID (Fréchet Inception Distance) [1] performance and parameter of state-of-the-art (SOTA) SR methods (lower FID values indicate better generative quality). EDSR [2], RCAN [3], and HAT-L [4] are regression-based models, typically generating low-quality distribution. The GAN-based approaches (MSRGAN [5], ESRGAN [6] and SPSR [7]) and DPM-based methods (SR3 [8] and IRSDE [9]) can produce high-quality images. Our EDiffSR achieves the best performance and is far more lightweight than SOTA DPM-based SR models.

Convolutional methods have shown significant success in modeling the non-linear relationship between LR and HR images in recent years. Among them, various efforts have been made to tame the inherent ill-posedness, such as dense [22], [23] and residual networks [24], attention-based models [3] [25] and transformer architectures [4] [26]. However, existing methods often employ regression function, *e.g.,* Mean Squared Error (MSE) and Mean Absolute Error (MAE), to minimize the pixel-level difference between super-resolved results and ground truth images. Despite obtaining favorable PSNR performance, these optimal objects can lead the model to average the pixel distance, resulting in over-smooth results. To restore visually convincing details, deep Generative Adversarial Networks (GANs) [27] have been explored. Such methods exploit the adversarial optimization between the generator and discriminator to encourage the generator to recover realistic images. Generally, GANs require carefully designed loss functions as auxiliary, *e.g.,* perceptual loss [6] and gradient loss [7], to optimize the distance in the feature domain. Although GANs can generate rich details, they often suffer from training instability and are easy to collapse, leading to undesirable artifacts.

Recently, Diffusion Probabilistic Models (DPM) [28] have received increasing attention in the realm of image-to-image translation, and also achieved promising performance in super-resolution tasks [29] [8] [30] [31]. The key to DPM is the reverse diffusion process, which iteratively predicts various noises from a noisy image. In this manner, DPM can generate high-quality data distributions from random noise. Thanks to its principled and well-defined probabilistic diffusion process, DPM can mitigate the training instability that commonly occurs in GANs and generate more complex distributions. More recently, Saharia *et al.* [8] pioneered the DDPM-based SR method, and utilized the heavy UNet as the denoiser to generate images by iterative refinement. To better simulate the degradation process, Luo *et al.* [9] proposed stochastic differential equations to model the diffusion process. Nevertheless, most DPM-based SR methods remain confined within the paradigm of image synthesis tasks, lacking insightful design for SR tasks. Specifically, **1) The prior knowledge in LR image, which is critical for SR tasks, is rarely explored.** Following the paradigm of image synthesis, the LR image is often directly upsampled by bicubic interpolation to serve as the condition. This pre-processing scheme lacks elaboration and can only convert partial prior knowledge into a diffusion model. As a result, it may lead to suboptimal performance. **2) The vanilla UNet consumes massive computational cost and is less effective in SR tasks.** In contrast to image synthesis which needs to predict an image from scratch, more pixels of SR are given. Thus, employing a large network for noise prediction is inefficient.

To this end, this paper explores the application of DPM and devises an Efficient Diffusion model for remote sensing image Super-Resolution (EDiffSR). Unlike previous work that applied bicubic-upsampled LR image as the condition, we developed a novel Conditional Prior Enhancement Module (CPEM) to effectively leverage prior knowledge in low-resolution (LR) images. It promotes the condition with more informative and plentiful input. Moreover, an Efficient Activation Block (EAB) is devised to form our denoising network (EANet), achieving favorable denoising capability while maintaining a far more lightweight design. As illustrated in Fig. 1, our EDiffSR achieves impressive performance while with significantly fewer parameters than previous DPM-based SR approaches (*e.g.,* SR3 [8] and IRSDE [9]). In addition, we equip our EDiffSR with the stochastic differential equations (SDE) [9] to further facilitate the sampling process in the diffusion process. Extensive evaluations on four remote sensing datasets demonstrate the superiority of our EDiffSR in both perceptual quality and quantitative metrics over the state-of-the-art CNN-based, GAN-based, and DPM-based models, while with considerable competitiveness in terms of model efficiency.

To sum up, the main contributions of this study are summarized as follows.

1) We pioneer an efficient yet effective diffusion probabilistic model (EDiffSR) for remote sensing image super-resolution. By introducing more prior knowledge into the diffusion model with elaborate CPEM, our EDiffSR can achieve accurate SR performance.
2) The proposed EANet can exceed the previous SOTA methods in noise prediction while with lower computational cost. It provides a new perspective for exploring more efficient diffusion-based frameworks.

The remainder of this paper is organized as follows. Section II reviews the progress of image super-resolution and diffusion models. Section III presents some preliminary of the diffusion process. Section IV introduces the implementation details of our EDiffSR. Section V contains experiments and analysis. Section VI is the conclusion.

II. Related Work

A. Deep Learning-based Image Super-Resolution

1) CNN-based Models: Inspired by the success of SRCNN [32], numerous elaborate CNN architectures have been proposed, such as very deep [24] and wide [2] architectures, attention mechanism [3] [25], and transformer [4]. Recently, Chen *et al.* [4] proposed an impressive method by combining the advantage of channel attention and self-attention to activate more useful information for SR. However, most CNN-based SR approaches predict the target distribution by minimizing the MSE (L_2) or MAE (L_1) loss. While achieving high PSNR values, the regression functions often tend to encourage the network to "average" a certain region, leading to an undesirable over-smooth issue. In contrast, our EDiffSR can benefit from the generative capability of DPM to recover more realistic distributions, improving the visual quality of SR results by a large margin.

2) GAN-based Models: To promote visual pleasure, GAN-based SR approaches introduce elaborate auxiliary loss to guide the network to generate photo-realistic results. For example, Ledig *et al.* [5] pioneered the perceptual loss that measures the feature-wise distance between the restored image and ground-truth image in VGG feature space. To improve the training stability, Wang *et al.* [6] put forward an Enhanced SRGAN (ESRGAN) with modified discriminative constraints while removing batch normalization (BN) to avoid artifacts. Sajjadi *et al.* [33] developed a texture loss to preserve the high-frequency textual details. Recently, Ma *et al.* [7] proposed a structural preserving GAN that maintains structural information by the gradient loss. Although GANs can bring impressive improvement in visual quality, they often face harsh optimization problems. Moreover, we often require laborious tricks to strike the balance between these carefully designed loss functions. Benefiting from the well-defined diffusion process, the proposed EDiffSR offers a stable and interpretable training process.

3) Diffusion-based Models: Diffusion models use a fixed Markov chain to optimize the variations boundaries of the likelihood function and have recently received increasing attention due to their excellent performance on generative tasks [34]. In SR task, research on diffusion modeling is still in its infancy. Until recently, Saharia *et al.* [8] proposed to generate results that exceed those of the GAN with iterative refinement. Li *et al.* [29] firstly introduced the residual prediction in DPM

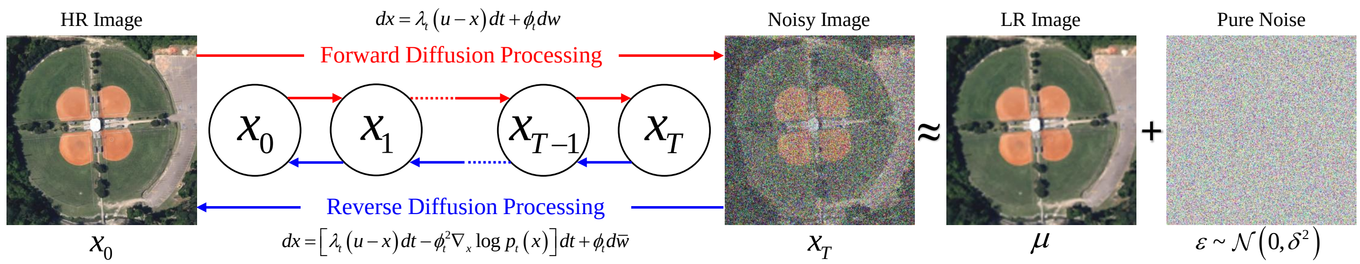

Fig. 2. Overview of the forward and reverse diffusion processes defined by mean-reverting Stochastic Differential Equations (SDE). The forward diffusion gradually degrades the high-quality and high-resolution image x_0 to its low-quality counterpart via $x_T = \mu + \varepsilon$. The reverse diffusion learns to characterize the noise and reconstruct the corresponding high-resolution image.

for face image SR. More recently, Xia *et al.* [35] exploited the transformer block to model the long-range discrepancy for effective image restoration. Luo *et al.* [9] proposed an averaging-equation idea to simulate the image degradation process while realizing a faster diffusion process. However, current DPM-based SR models often rely on large models for noise prediction. The high complexity of denoisers limits their practical application and leads to inefficient inference in large-scale remote sensing scenarios. In contrast, the proposed EDiffSR achieves favorable noise prediction with a far more lightweight EANet. Besides, existing methods barely consider the prior information in images, which is crucial for SR tasks, thus resulting in suboptimal performance. The proposed EDiffSR seeks a more desirable condition by exploring informative cues from the LR images, which further boosts the SR performance.

B. Remote Sensing Image Super-Resoltuion

Early SR methods are CNN-based, aiming to achieve high PSNR performance [36] [37] [38]. In this parse, more efforts have been paid to improve the network structure, making the convolution network grasp more characteristics of remote sensing images, such as scene-adaptive network [39], multi-scale [40] [41] [42] and multi-stage [37] design, numerous attentions [43] [44] and guided super-resolution [45]. Li *et al.* [46] proposed a novel dual-stage network to reconstruct more missing details in remote sensing imagery in a coarse-to-fine manner. Recently, Li *et al.* [47] put forward to transfer more beneficial supplementary from RGB images to remote sensing scenes. However, as PSNR tends to penalize the reconstruction of high-frequency details, these methods can not reflect human preference well in RSI.

To recover rich detailed information in RSI, various GAN-based methods have been proposed. Lei et al. [48] a coupled-discriminated GAN to make better discrimination. Jiang *et al.* [49] proposed an edge-enhanced GAN by optimizing the high-frequency and low-frequency components simultaneously. Haut *et al.* [50] proposed to train a GAN in an unsupervised manner without the HR RSI. Recently, Tu *et al.* [51] incorporated the long-range modeling capability of the Swin transformer into GAN, achieving favorable perceptual quality of SR results. However, these approaches often involve complex optimization functions and network structures, leading to training instability. In contrast, this paper proposed an efficient solution to recover perceptual-pleasant RSI, mitigating the training instability of GAN.

In the area of remote sensing, some researchers have applied diffusion modeling to SR tasks [52] [53] [30]. However, they borrow too much from the paradigm in image synthesis, which uses a large UNet for noise estimation, resulting in inefficient inference in SR tasks. In addition, there is a lack of consideration of incorporating the valuable prior knowledge in diffusion to generate high-frequency details in remote sensing images. In this paper, we demonstrate that a low-complexity network can provide a more practical and efficient scheme to deliver competitive denoising performance in SR tasks when compared to state-of-the-art methods employing larger models like UNet. In addition, unlike simple bicubic upsampling, we choose to explore more prior information to generate informative conditions, thus further enhancing the diffusion model to generate realistic distributions.

III. Preliminary

A. Forward Diffusion Process

The forward diffusion process aims to gradually transform the initial data distribution x_0 to a noisy image x_T after time step T. As shown in Fig. 2, we define the ground truth image I_{HR} as x_0. As such, x_T can approximately to the combination of the bicubic-upsampled LR image μ and a pure Gaussian noise $\varepsilon \sim \mathcal{N}(0, \delta^2)$. Here δ^2 represents the stationary variance. This paper adopts the mean-reverting Stochastic Differential Equations (SDE) [9] to define the diffusion process as it allows for an efficient sampling process. Specifically, as illustrated in Fig. 2, the forward diffusion process is depicted as

$$dx = \lambda_t \left(u - x\right) dt + \phi_t dw, \tag{1}$$

where w refers to a standard Wiener process. λ_t and ρ_t are two time-dependent parameters that control the speed of mean reversion and stochastic volatility, respectively. To make equation 1 have a closed-form solution, we set $\phi_t^2/\lambda_t = 2\delta^2$. As shown in Fig. 2, given an HR image x_0 and $t \in [0, T]$, for an intermediate moment t, the corresponding state x_t can be strictly expressed by the closed-form solution of Eq. (1):

$$x_t = \mu + (x_0 - \mu)e^{-\bar{\lambda}_t} + \int_0^t \phi_z e^{-\bar{\lambda}_t} dw(z), \tag{2}$$

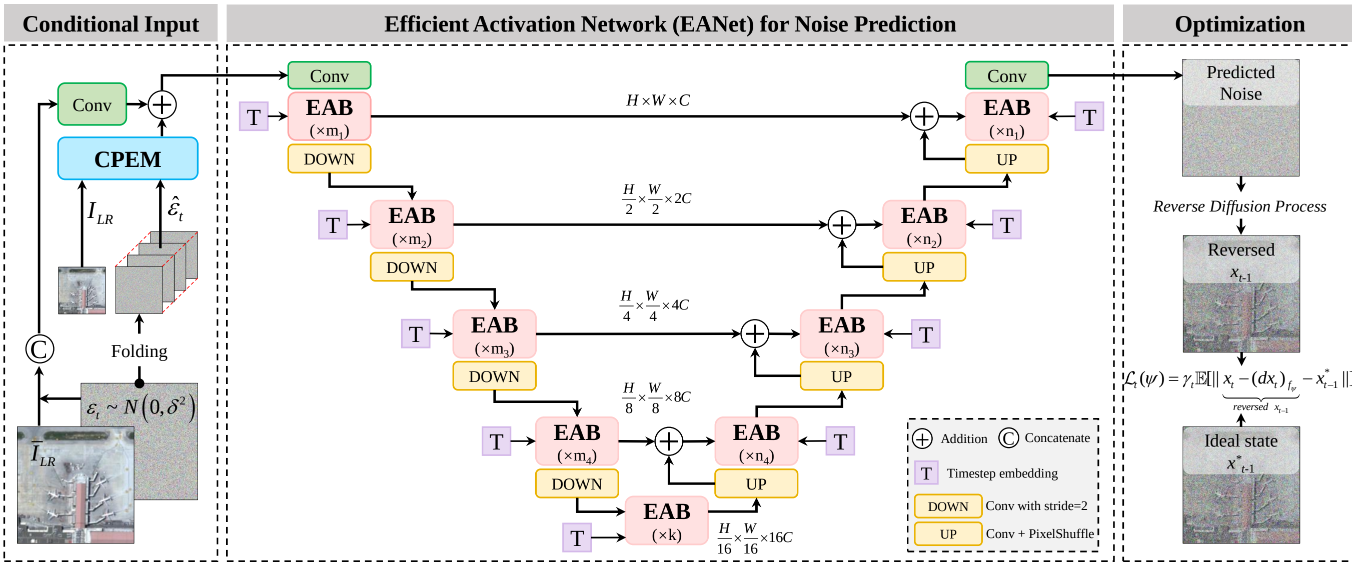

Fig. 3. Overall framework of our EDiffSR. It consists of three parts: the condition part, Efficient Activation Network (EANet), and the optimization part. In the condition part, CPEM is designed to explore more priors from the original LR image. EANet takes the condition as input and characterizes the noise distribution. Compared to the primary Unet, it is more efficient and effective owing to the efficient activation block (EAB). The optimization process adopts the maximum likelihood learning for a more stable diffusion process.

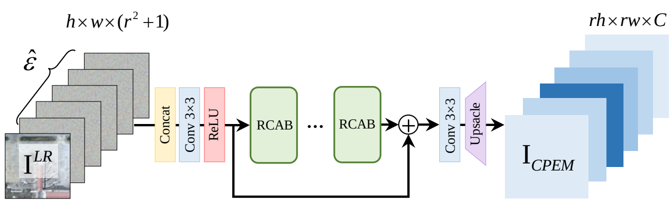

Fig. 4. The flowchart of Conditional Prior Enhancement Module (CPEM). We adopt the Residual Channel Attention Block (RCAN) to extract rich prior information.

where $\bar{\lambda}_t$ is equal to $\int_0^t \lambda_z dz$. The proof of Eq. 2 can be found at [9]. In this case, x_t follows a Gaussian probability distribution $p_t(x)$, expressed as

$$x_t \sim p_t(x) = \mathcal{N}(x_t|m_t(x), n_t) \tag{3}$$

where $m_t(x) = \mu + (x_0 - \mu)e^{-\bar{\lambda}_t}$ and $n_t = \delta^2(1 - e^{-2\bar{\lambda}_t})$ are the mean and variance of this Gaussian distribution, respectively. It is easy to observe that as the diffusion time $t \to \infty$, m_t and n_t would converge to μ and δ^2, *i.e.,* the terminal state $x_T \approx \mu + \varepsilon$, which aligns with the aim of the forward diffusion process.

B. Reverse Diffusion Process

Reverse diffusion aims to recover the HR image from the terminal state x_T. We can define the reverse diffusion process by simulating the reverse-time SDE [54] as

$$dx = \left[\lambda_t (u - x)\, dt - \phi_t^2 \nabla_x \log p_t(x)\right] dt + \phi_t d\bar{w}. \tag{4}$$

where $\bar{w}$ denotes a reverse-time Wiener process. $\nabla_x \log p_t(x)$ is the ground-truth score during inference stage. Note that in the training stage, the ground-truth image x_0 is available, thus we can leverage more pleasurable conditional scores during model training. In particular, it can be defined by

$$\nabla_x \log p_t(x|x_0) = -\frac{x_t - m_t(x)}{n_t}. \tag{5}$$

Furthermore, if we re-parameterize x_t to $x_t = m_t(x) + \sqrt{n_t}\varepsilon_t$, where ε_t is a standard Gaussian noise with the distribution $\mathcal{N}(0, I)$. The ground-truth scores can be expressed as $-\frac{\varepsilon_t}{\sqrt{n_t}}$. Since $m_t(x)$ and n_t are known, then we just need to estimate the noise using a noise prediction network f_ψ.

Similar to DDPM [28], we compute the Euclidean distance between the predicted noise and ground truth noise ε_t by the following formula:

$$\mathcal{L}(\psi) = \sum_{t=0}^{T} \gamma_t \mathbb{E}[|| \underbrace{f_\psi(x_t, u, v, t)}_{predicted\ noise\ \bar{\varepsilon}_t} - \varepsilon_t||], \tag{6}$$

where γ_t denotes the positive weight, and v refers to the original LR image.

IV. Proposed Method

A. Overview

Fig. 3 details the flowchart of our proposed EDiffSR. In the input part, we perform conditional prior enhancement to generate a more pleasurable condition for noise prediction. Specifically, the prior enhancement module f_{CPEM} takes the random noise ε_t, LR image v, and the corresponding bicubic-upsampled LR image $\bar{I}_{LR}$ as input, and then produce the enriched condition I_t by the following formula:

$$\mathrm{I}_t = f_{CPEM}(v, \hat{\varepsilon}_t) + \mathrm{f}_3([\mu, \varepsilon_t]), \tag{7}$$

where $\hat{\varepsilon}_t = \mathrm{Fold}(\varepsilon_t)$ represents that we adopt the pixel-folding operator to downsample the scale of ε_t without loss spatial information. $\mathrm{f}_3(\cdot)$ is a 3×3 convolution, and $[\cdot]$ represents channel-wise concatenation. Subsequently, a conditional time-dependent network f_ψ takes the pleasurable condition and time t as input, aiming to output a pure noise:

$$\bar{\varepsilon}_t = f_\psi(\mathrm{I}_t, t)\,. \tag{8}$$

Here, we adopt the efficient activation network (EANet) for noise prediction. Finally, we can optimize f_ψ until it converges.

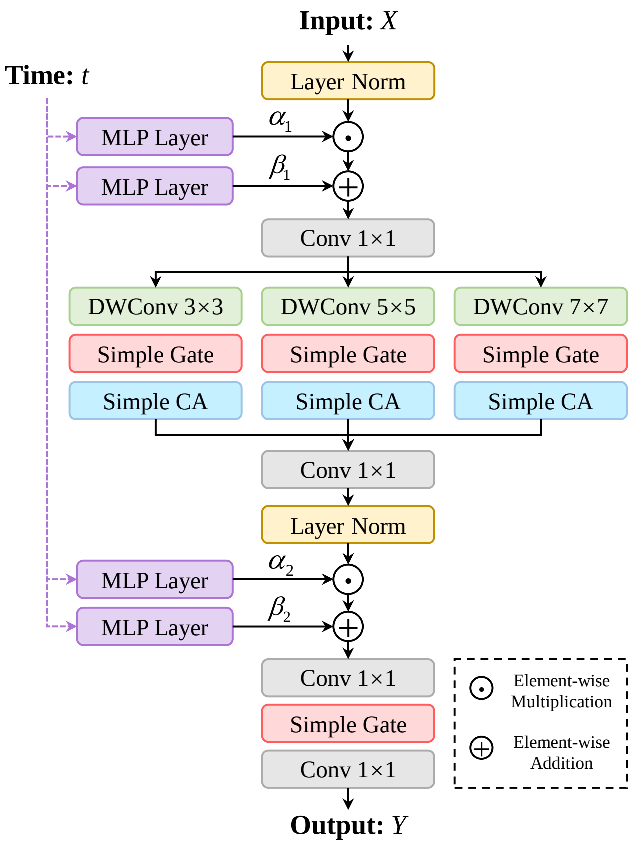

Fig. 5. Illustration of the Efficient Activation Block (EAB).

B. Conditional Prior Enhancement Module

Most previous methods have typically prepared the condition input by simply upsampling the LR image using bicubic interpolation. However, for SR tasks, this scheme may lose critical structure information, resulting in suboptimal conditional inputs. In contrast, our EDiffSR proposes to generate a more informative condition by exploring additional prior knowledge from the LR image, thus enriching the condition information for better SR performance.

As shown in Fig. 4, the CPEM mainly consists of a convolution layer and ReLU activation, followed by stacked Residual Channel Attention Blocks (RCABs) [3] and an upscale layer. To unify the scale of noise and the LR image, we first convert ε_t to the noisy cube using pixel folding. Then we concatenate and pass them through a 3×3 convolution layer followed by ReLU activation to perform shallow feature extraction, depicted as

$$\mathrm{I}_0 = \mathrm{ReLU}(\mathrm{Conv}(\mathrm{Concat}(v, \hat{\varepsilon}_t))). \tag{9}$$

Subsequently, n cascaded RCABs f_{RCAB} are used to achieve deep feature extraction and stabilize the gradient by global residual connection.

$$\mathrm{I}_{deep} = f^{n}_{RCAN}(\mathrm{I}_0) + \mathrm{I}_0. \tag{10}$$

Following that, a 3×3 convolution and a PixelShuffle layer [55] are used to get the condition yield from CPEM.

$$\mathrm{I}_{CPEM} = \mathrm{PixelShuffle}(\mathrm{Conv}(\mathrm{I}_{deep})). \tag{11}$$

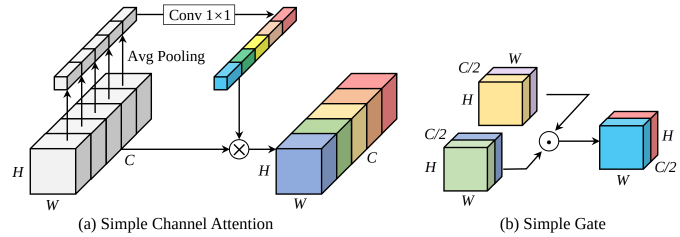

Fig. 6. The flowchart of (a) Simple Channel Attention and (b) Simple Gate operation.

According to Eq. (7), we generate the improved condition.

C. Efficient Activation Network for Noise Prediction

As illustrated in Fig. 3, the key component of our EANet is the Efficient Activation Block (EAB). Fig. 5 displays the architecture of the EAB, showcasing its lightweight design. The EAB primarily consists of Depth-Wise Convolution (DWConv), simple channel attention, and simple gate operations. This lightweight design results in significantly lower computational complexity when compared to large UNet architectures that incorporate channel attention or self-attention mechanisms. As discussed before, in the context of SR tasks, the majority of pixels are known. Therefore, a large model running massive calculations is inefficient in SR and may lead to a suboptimal performance due to redundant inference. Our EANet offers a more practical scheme to achieve favorable denoising performance with a lightweight model.

As shown in Fig. 5, given an input X and time step t, EAB predict an output Y. In particular, t will be projected to two flatten features by MLP layer for feature modulation:

$$\mathrm{F} = f_{1\times1}(\alpha_1 \odot \mathrm{Norm}(\mathrm{X}) + \beta_1). \tag{12}$$

Subsequently, we use multi-scale DWConv to explore the multi-scale knowledge in RSI. Within each scale, we incorporate additional nonlinear representations through the use of simple channel attention and simple gate activation, as illustrated in Fig. 6. To simplify the channel attention [3], we eliminate the convolution layer and sigmoid activation. The simple gate activation is essentially an element-wise product operation applied to the feature maps. The multi-scale simple activation process can be expressed as

$$\begin{cases} \mathrm{F}^3 = \mathrm{SCA}(\mathrm{SimpleGate}(f_{3\times3}(\mathrm{F}))) \\ \mathrm{F}^5 = \mathrm{SCA}(\mathrm{SimpleGate}(f_{5\times5}(\mathrm{F}))) \\ \mathrm{F}^7 = \mathrm{SCA}(\mathrm{SimpleGate}(f_{7\times7}(\mathrm{F}))) \end{cases} \tag{13}$$

After that, we set a 1×1 convolution to aggregate these multi-scale representations $\mathrm{F}' = f_{1\times1}(\mathrm{Concat}(\mathrm{F}^3, \mathrm{F}^5, \mathrm{F}^7))$. After layer norm, we conduct modulation with the scaling and the shifting operation:

$$\bar{\mathrm{F}} = \alpha_2 \odot \mathrm{Norm}(\mathrm{F}') + \beta_2. \tag{14}$$

At the final, the output Y can be obtained via the following formulation.

$$\mathrm{Y} = f_{1\times1}(\mathrm{SimpleGate}(f_{1\times1}(\bar{\mathrm{F}})). \tag{15}$$

Following previous works [9], [30], we form our EANet to a U-shape encoder-decoder structure. During the encoding

Algorithm 1: Training of our EDiffSR

Input: HR image $x_0 = I_{HR}$,LR image $v = I_{LR}$, upsampled LR image $\mu = \bar{I}_{LR}$, total step T.

Initialization: Random sample $\varepsilon_t \sim \mathcal{N}(0, \delta^2)$, $t \in [0, T]$, $T = 100$.

repeat

$\mathrm{I}_t = f_{CPEM}(v, \hat{\varepsilon}_t) + \mathrm{Conv}([\mu, \varepsilon_t])$; `// Enhance`

$\bar{\varepsilon}_t = f_\psi(\mathrm{I}_t, t)$; `// Predict noise`

`// Substitute score into (6)`

$dx_t = [\lambda_t (u - x_t) dt - \phi_t^2 \frac{\bar{\varepsilon}_t}{\sqrt{n_t}}]dt + \phi_t d\bar{w}$;

$\mathcal{L}(\psi) = \gamma_t \mathbb{E}[|| \underbrace{x_t - (dx_t)_{f_\psi}}_{reversed\ x_{t-1}} - x_{t-1}^* ||]$; `// Loss`

$\nabla_\psi \mathcal{L}$; `// Gradient descent`

until converged

Algorithm 2: Inference of our EDiffSR

Input: LR image $v = I_{LR}$, upsampled LR image $\mu = \bar{I}_{LR}$, total step T.

Output: The super-resolved image I_{SR}.

Initialization: Random sample $x_T \sim \mathcal{N}(0, \delta^2)$, f_ψ is the pre-trained EANet, $\mathrm{EM}(\cdot)$ is Euler–Maruyama method, $T = 100$.

for $t = T : 1$ **do**

$\bar{\varepsilon}_t = f_\psi(x_t, u, v, t)$; `// Predict noise`

`// Substitute score into (6)`

$dx_t = [\lambda_t (u - x_t) dt - \phi_t^2 \frac{\bar{\varepsilon}_t}{\sqrt{n_t}}]dt + \phi_t d\bar{w}$;

$x_{t-1} = x_t - \mathrm{EM}(dx_t)$; `// Reverse SDE`

end

$I_{SR} = x_0$;

phase, we employ a sequence of EABs and a convolution operation with a stride of 2 to progressively downsample the feature maps. In the decoding phase, multiple EABs and pixel-shuffle layers are used to upscale the features. The number of EABs in the encoder and decoder component is denoted as $[m_1, m_2, m_3, m_4]$ and $[n_1, n_2, n_3, n_4]$, respectively. Additionally, we introduce k EABs in the middle of the EANet.

D. Optimization and Inference

Although Eq. (6) can provide a straightforward solution to optimize the EANet, the diffusion model often suffers from instability in the training process. The key reason is predicting an instantaneous distribution of noise is not an easy task. Therefore, we modified the training object by using a maximum likelihood learning strategy used in [9]. To optimize EANet, specifically, we choose to minimize the Euclidean distance below:

$$\mathcal{L}(\psi) = \sum\nolimits_{t=0}^{T} \gamma_t \mathbb{E}[|| \underbrace{x_t - (dx_t)_{f_\psi}}_{reversed\ x_{t-1}} - x_{t-1}^* ||], \tag{16}$$

where x_{t-1}^* is the ideal state reversed from x_t. The closed-form of x_{t-1}^* can be determined by the following formula:

$$\begin{aligned} x_{t-1}^* =& \frac{1 - e^{-2\bar{\lambda}_{t-1}}}{1 - e^{-2\bar{\lambda}_t}} e^{-\lambda'_t} (x_t - \mu) \\ &+ \frac{1 - e^{-2\lambda'_t}}{1 - e^{-2\bar{\lambda}_t}} e^{-\bar{\lambda}_{t-1}} (x_0 - \mu) + \mu. \end{aligned} \tag{17}$$

The proof can be referred to [9]. In brief, we transformed the distance between the predicted noise and ground-truth noise into another domain, *i.e.,* the distance between the ideal state and predicted states. This scheme helps to reduce the optimization instability as most pixels in reversed states are known.

In the inference procedure, we utilize the pre-trained f_ψ to predict high-resolution images by sampling from the random state x_T and iteratively solve the SDE with numerical solutions, such as the Euler–Maruyama method [56]. To better understand the training and inference process of our EDiffSR, we summarize these processes in two algorithms, as presented in Algorithm. 1 and Algorithm 2.

V. Experiment and Discussion

In this section, we conduct extensive experiments on four remote-sensing datasets to evaluate the performance of our EDiffSR, both in simulated and real-world scenarios.

A. Dataset

We use 4 public remote sensing datasets to comprehensively evaluate the effectiveness of the methods in this paper, including AID [57], DOTA [58], DIOR [59], and NWPU-RESISC45 [60]. The training set in this study consists of 3,000 randomly selected images from the AID dataset with an image size of 640×640. Specifically, we randomly select 100 images in each of the 30 categories of AID to build the training set. Additionally, we have selected 10 images from each category that do not overlap with the training set to form the test set, resulting in a total of 300 test images. Furthermore, we have used a subset of images from the DOTA dataset and the DIOR dataset for testing, consisting of 700 and 1,000 images, respectively. These images have a resolution of 512×512. As a result, our test set comprises a total of 2,000 images. In our simulated experiments, we used bicubic interpolation for image degradation. NWPU-RESISC45 data were only used for real-world analysis without any simulated degradation. To save the inference cost, we randomly selected 315 images from NWPU and cropped them to 128×128.

B. Implementation Details

This study focus on $\times 4$ SR, *i.e.,* $r = 4$. In our final EDiffSR, we incorporate 5 RCAB in the CPEM for prior enhancement while maintaining a favorable model size. The inner-channel number in EANet is set to $C = 64$. Following prior works [30] [9], the depth of the noise prediction network is set to 4. In particular, the number of EAB in each depth $[m_1, m_2, m_3, m_4]$ and $[n_1, n_2, n_3, n_4]$ are set to $[14, 1, 1, 1]$ and $[1, 1, 1, 1]$, respectively. We include one EAB at the middle layer of EANet, *i.e.,* $k = 1$. To train our EDiffSR, we perform

TABLE I. Quantitative FID comparison with state-of-the-art SR models on 30 scene categories of the AID test set. The best FID value in each category is highlighted in **red** while the second best is in **blue**.

Categories	Bicubic	EDSR [2]	RCAN [3]	HAT-L [4]	MSRGAN [5]	ESRGAN [6]	SPSR [7]	SR3 [8]	IRSDE [9]	EDiffSR
Airport	126.23	87.25	89.52	86.55	54.42	54.85	57.98	56.56	**54.27**	**52.76**
Bare Land	113.49	91.50	92.51	91.15	66.83	**60.75**	72.17	76.89	80.30	**66.76**
Baseball Field	131.28	89.17	91.09	90.06	**51.25**	**46.87**	55.88	70.22	57.67	52.43
Beach	121.31	104.78	106.03	101.27	50.90	48.96	52.78	50.81	**43.22**	**43.10**
Bridge	137.67	80.01	82.27	80.93	**47.43**	50.70	49.40	73.65	**45.71**	50.98
Center	140.22	71.42	74.34	71.55	50.27	54.30	48.24	52.00	**44.29**	**43.13**
Church	122.26	85.76	87.92	89.48	51.88	51.89	55.03	62.58	**50.76**	**50.70**
Commercial	112.45	109.99	110.88	104.21	55.42	56.18	60.77	69.68	**51.84**	**55.20**
DenseResidential	126.36	113.85	125.26	118.28	52.16	57.75	55.99	62.96	**39.53**	**39.97**
Desert	115.10	77.50	76.95	75.84	56.28	**53.54**	64.02	59.35	61.50	**52.91**
Farmland	144.78	92.15	93.80	96.13	66.34	**56.12**	58.00	79.39	61.85	**51.07**
Forest	103.57	88.79	93.45	96.26	59.69	64.36	62.01	72.07	**48.68**	**46.90**
Industrial	106.82	77.84	80.85	74.83	37.79	**37.11**	45.90	46.07	**35.90**	41.27
Meadow	133.81	107.78	106.18	103.1	95.57	68.95	**64.83**	87.56	70.93	**66.53**
MediumResidential	117.19	98.74	104.17	100.24	46.17	50.11	49.75	73.45	**41.80**	**40.05**
Mountain	103.15	105.5	105.64	103.68	57.93	**54.93**	71.01	72.67	59.02	**52.21**
Park	137.79	109.64	112.28	109.54	**61.02**	**60.33**	72.14	80.81	63.78	63.29
Parking	134.86	60.79	67.40	63.98	41.79	42.99	45.40	56.09	**37.02**	**36.74**
Playground	113.86	58.07	61.90	60.36	40.69	39.15	41.00	53.96	**38.89**	**35.94**
Pond	162.50	122.29	124.27	126.31	60.65	**54.71**	62.62	104.36	56.29	**55.97**
Port	134.94	77.12	80.02	80.55	46.76	**46.72**	52.02	58.93	**46.52**	48.22
Railway Station	113.35	93.26	93.77	87.06	50.38	52.08	58.44	56.59	**49.82**	**51.89**
Resort	131.05	99.11	104.87	105.46	59.79	61.77	67.71	69.35	**59.00**	**57.26**
River	151.14	106.24	109.20	108.06	**54.50**	59.23	65.18	83.28	60.27	**57.34**
School	110.22	85.48	89.16	82.25	49.53	50.33	53.65	60.20	**47.60**	**47.00**
SparseResidential	149.02	134.24	140.41	132.73	73.57	75.55	77.83	85.06	**69.59**	**71.52**
Square	108.42	70.79	75.52	71.89	**42.48**	44.89	46.29	52.92	45.04	**42.43**
Stadium	121.79	56.48	59.39	58.87	37.70	35.28	37.42	38.27	**32.59**	**34.18**
Storage Tanks	161.44	89.80	93.90	88.43	45.57	51.67	51.01	53.38	**45.09**	**42.93**
Viaduct	109.83	66.8	68.89	66.65	36.11	34.32	37.66	46.14	**33.55**	**32.81**
Average	126.53	90.40	93.39	90.86	53.36	52.55	56.40	65.51	**51.08**	**49.45**

500,000 iterations with a mini-batch size of 4. The initial learning rate is set to 4×10^{-5} and decays following a cosine schedule. We utilize the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The total step of the diffusion process is $T = 100$. All SR methods involved in this paper were retrained from scratch on the AID training set. For a fair comparison, we did not perform any pre-training and fine-tuning processes in our EDiffSR. Our experiments are implemented on PyTorch with a 24GB memory NVIDIA RTX 3090 GPU and a 3.40 GHz AMD Ryzen 5700X CPU.

C. Metrics

In this paper, 7 metrics are used to comprehensively evaluate the performance of SR model. In the simulation experiments, where the ground-truth image is available, we utilize 5 full-reference metrics: FID (Fréchet Inception Distance) [1], LPIPS (Learned Perceptual Image Patch Similarity) [61], DISTS (Deep Image Structure and Texture Similarity) [62], as well as the widely used PSNR (Peak Signal-to-Noise Ratio) and the SSIM (Structural Similarity Index) [63]. These metrics help assess the distance between the generated images and the ground truth images. Among them, FID is widely used to measure the generative quality of the generative model. It enhances the Inception Score (IS) [64] metric by directly measuring the feature-level distance without the need for a classifier. In real-world experiments without ground-truth images, we additionally report the results on two reference-free metrics: NIQE (Natural Image Quality Evaluator) [65] and AG (Average Gradient). These metrics offer insights into the perceptual quality and the high-frequency details of the generated images.

D. Comparison With State-of-the-Arts

We compared our EDiffSR with state-of-the-art (SOTA) SR approaches, including EDSR [2], RCAN [3], HAT-L [4], MSRGAN [5], ESRGAN [6], SPSR [7], SR3 [8], IRSDE [9]. We selected these methods as they represent the mainstream approaches in the field, ensuring a comprehensive evaluation. Specifically, EDSR, RCAN, and HAT-L are CNN-based approaches that adopt wide CNN, channel attention, and transformer architectures, respectively. Note that MSRGAN is a modified version of SRGAN, where the BN layer is removed to avoid artifacts, and it employs the same perceptual loss as ESRGAN. SPSR employs carefully designed gradient loss to preserve structural details and has demonstrated favorable performance. On the other hand, SR3 and IRSDE are SOTA diffusion-based models. We retrained these comparative approaches on the AID training set according to their official implementation settings.

1) Quantitative Comparison: Results of FID values on 30 categories of the AID test set are reported in Table. I. In each row, we highlighted the best and the second-best FID performance. We can find in most remote sensing scenes our EDiffSR achieves favorable FID performance against all comparative models. However, due to the complex diversity of remote sensing scenes, achieving generalization across various

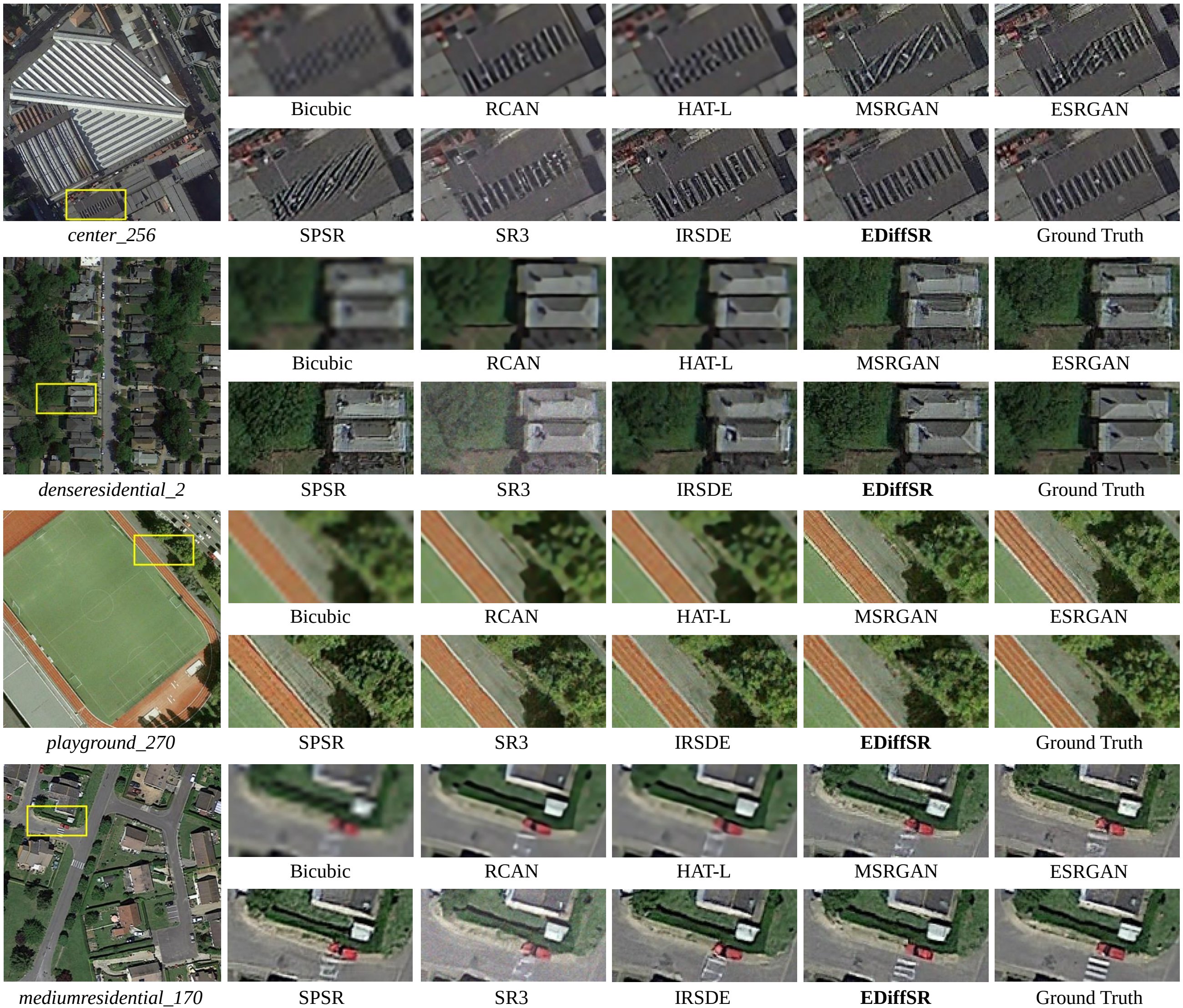

Fig. 7. $\times 4$ visual comparisons with state-of-the-art SR models on AID test set. The results show that our EDiffSR significantly outperforms comparative approaches in high-frequency detail recovery while producing visually pleasing images that are more natural. Zoom in for a better view.

scenes remains a challenging task. Specifically, EDiffSR outperforms the second-best approach (IRSDE) by an average margin of 1.63 in terms of FID. These results reveal that EDiffSR can provide robust high-quality data distribution in various remote sensing scenarios, highlighting its favorable generative capability.

Additionally, we presented the average FID, LPIPS, DISTS, and NIQE values across the AID, DOTA, and DIOR test sets in Table II. We observed that our EDiffSR still achieves the best FID performance across these test sets. It is worth highlighting that GAN-based models excel in achieving the best LPIPS results because they usually adopt the VGG space to compute the perceptual loss, which aligns with the calculation of LPIPS. In this case, our EDiffSR achieves acceptable LPIPS results and surpasses Diffusion-based approaches by a large margin. For instance, compared to SR3, EDiffSR exhibits a remarkable 0.0647 improvement in terms of LPIPS. When compared to IRSDE, we achieve superior LPIPS performance (0.1898 vs. 0.2419) in the DIOR dataset. Notably, both IRSDE and EDiffSR utilize the same diffusion process equation, *i.e.,* SDE. Therefore, the results demonstrate the superiority of our EANet in providing effective noise prediction capabilities compared to the commonly employed UNet architecture in IRSDE. As for the DISTS metrics, we observe that despite SPSR focusing on preserving structural details, it only achieves the second-best performance on the DOTA and DIOR datasets. In contrast, our EDiffSR excels in attaining the best DISTS scores for both DOTA and DIOR, highlighting its remarkable capability to restore accurate structural details in remote sensing images. Moreover, EDiffSR can achieve the best NIQE values in almost all test sets. As a result, the proposed EDiffSR does recover realistic results that align well with human perception.

Besides the above results, the PSNR and SSIM results are also tabulated in Table. IV. The best and second-best performance within each category of methods are highlighted in bold and underlined. PSNR-oriented models such as EDSR, RCAN, and HAT-L achieve higher PSNR and SSIM scores

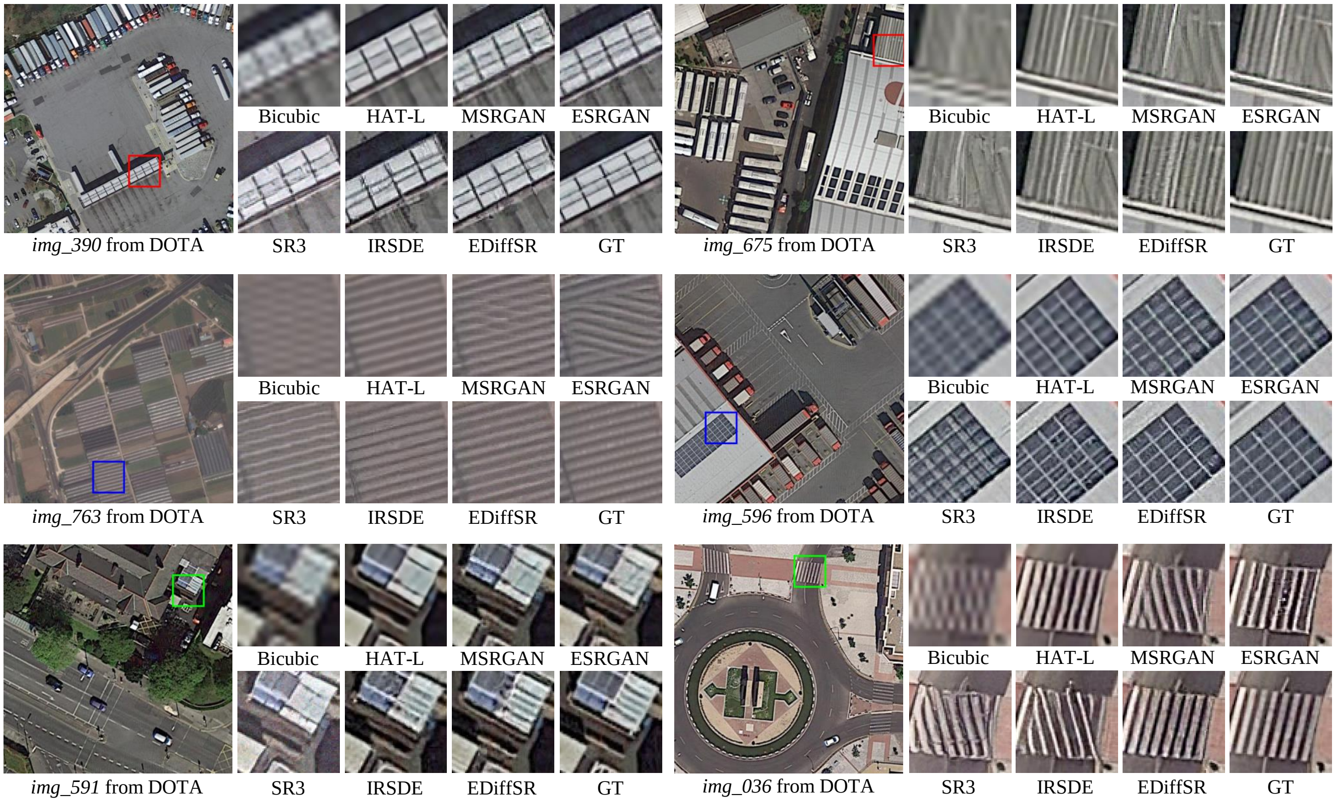

Fig. 8. ×4 visual comparisons with state-of-the-art SR models on DOTA test set. Zoom in for a better view.

when compared to GAN-based and DPM-based SR methods. This is because they optimize Mean Squared Error (MSE), which provides a straightforward learning objective for high PSNR performance. In particular, PSNR and SSIM are inconsistent with human perceptual while guiding the network to generate over-smooth content. As demonstrated in Table. II, despite HAT-L achieving the highest PSNR score, it gains undesired scores in terms of FID, DISTS, and LPIPS. Furthermore, PSNR-driven methods tend to produce blurry results, resulting in poor perceptual quality. In the DPM-based category, our EDiffSR consistently delivers higher-quality results while maintaining the best PSNR/SSIM performance. When compared to SR3, we achieve a significant improvement in terms of PSNR (27.40dB vs. 26.24dB) in the AID test set, demonstrating that our lightweight EANet is capable of providing excellent denoising performance in SR tasks.

2) Qualitative Comparison: We conducted a visual comparison with all comparative models. From Fig. 7, we find that our EDiffSR can consistently produce photo-realistic results that surpass SOTA approaches. For the ”center_256” in AID, both RCAN and HAT-L produce blurry results, highlighting the limited generalization of PSNR-oriented models in recovering rich details. In contrast, GAN-based models can restore shape details, especially edge information, but often introduce severe artifacts inconsistent with the ground truth. EDiffSR consistently delivers more natural and realistic results, demonstrating its capacity to generate visually pleasing images. For the ”mediumresidential_170” image, RCAN and HAT-L still exhibit an over-smoothed appearance, while other GAN-based and diffusion-based models yield more natural results with realistic details. Compared with IRSDE and SR3 which directly adopt the bicubic interpolation to prepare conditions, EdiffSR contains more context details that are close to the ground-truth image, such as the the marks on the road. The results demonstrate that the proposed CPEM is helpful in exploring more useful priors, *e.g.,* edge information, thus boosting the performance of DPM in SR tasks.

In Fig. 8, we also visualize some SR results on the DOTA test set. As shown in Fig. 8, both CNN-based approaches and GAN-based models fail to achieve satisfactory detail, especially in terms of edges and textures. For ”img_591” from DOTA, only EDiffSR successfully restores the clear and sharp details of the building on the ground. In ”img_036”, MSRGAN, SR3, and IRSDE exhibit severe distortion, which deviates from the ground-truth distribution. HAT-L and ESRGAN can offer relatively realistic distribution, restoring accurate direction of the lines on the road. Nevertheless, the results obtained from ESRGAN exhibit an unnatural appearance due to the oversharpening issue. In contrast, EDiffSR accurately generates these details and appears more natural perception. These results highlight the capability of CPEM to explore additional prior information, enabling EDiffSR to recover more details that align with the realistic distribution of the ground truth.

In Fig. 9, we zoomed in and displayed some visual results from the DOIR dataset. As shown in ”img_895”, EDiffSR exhibits an impressive visual performance, outperforming other methods in accurately recovering the direction of the lines

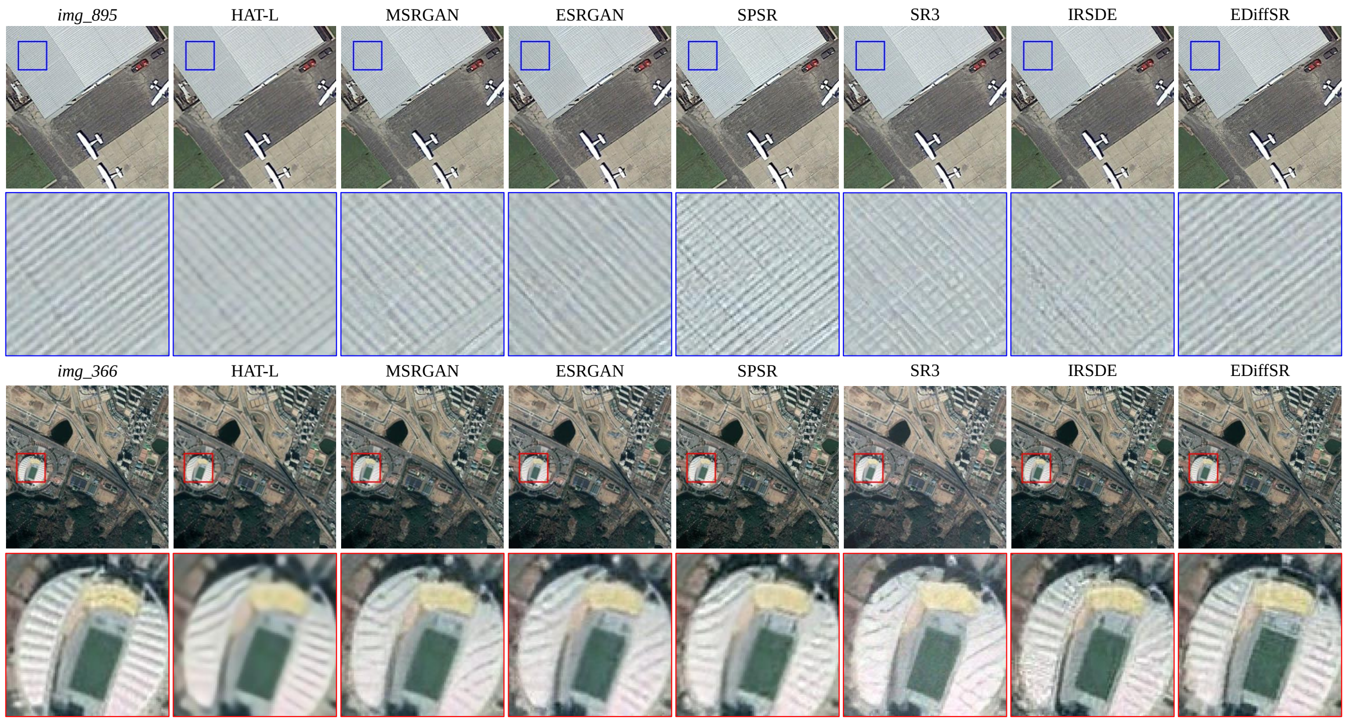

Fig. 9. $\times 4$ visual comparisons with state-of-the-art SR models on DIOR test set. Zoom in for a better view.

TABLE II. Quantitative comparison with state-of-the-art SR models in terms of FID, LPIPS, DISTS, and NIQE across AID, DOTA, and DIOR test sets. The best performance value is highlighted in **red** while the second best is in **blue**.

Dataset	Metrics	Baseline	CNN-based			GAN-based			Diffusion-based		
		Bicubic	EDSR [2]	RCAN [3]	HAT-L [4]	MSRGAN [6]	ESRGAN [6]	SPSR [7]	SR3 [8]	IRSDE [9]	EDiffSR
AID [57]	FID ↓	126.53	90.40	93.39	90.86	56.36	52.55	56.40	65.51	**51.08**	**49.45**
	LPIPS ↓	0.4801	0.3068	0.3112	0.3078	**0.1694**	**0.1695**	0.1751	0.2534	0.2247	0.1887
	DISTS ↓	0.1512	0.0880	0.0900	0.0882	0.0590	0.0601	0.0632	0.0903	**0.0579**	**0.0561**
	NIQE ↓	21.24	19.79	19.88	20.15	17.46	15.51	17.90	15.16	**14.80**	**14.22**
DOTA [2]	FID ↓	65.99	55.22	43.37	42.61	25.12	24.42	26.42	35.74	**23.91**	**21.26**
	LPIPS ↓	0.4416	0.2616	0.2629	0.2641	0.1649	**0.1506**	**0.1605**	0.2790	0.1919	0.1689
	DISTS ↓	0.1481	0.0831	0.0846	0.0832	0.0589	0.0582	0.0617	0.0993	**0.0550**	**0.0556**
	NIQE ↓	19.48	18.37	18.58	18.98	16.71	15.48	18.01	15.49	**14.47**	**14.20**
DIOR [59]	FID ↓	57.42	41.87	43.20	41.94	23.16	22.51	24.02	29.60	**22.28**	**21.79**
	LPIPS ↓	0.4678	0.3020	0.3048	0.3062	**0.1722**	0.1836	**0.1772**	0.2836	0.2419	0.1898
	DISTS ↓	0.1497	0.0886	0.0899	0.0893	**0.0605**	0.0623	0.0654	0.0924	0.0622	**0.0590**
	NIQE ↓	20.24	19.16	19.30	19.56	17.61	**15.13**	18.24	15.55	15.19	**15.16**

on the building roof. In this context, reconstructing such high-frequency information can be challenging. All methods yielded a completely wrong distribution of these details, except our EDiffSR. Benefiting from our condition prior enhancement module (CPEM), more high-frequency prior information can be explored and introduced into the diffusion process, making the output images consistent with the spatial distribution of ground-truth images. In "img_366", we zoomed in on the stadium region for comparison. It is evident that our EDiffSR reconstructs the most realistic results, whereas the other models exhibit significant distortion and blurring.

3) Real-world Comparison: We also evaluate the performance of our EDiffSR on real-world remote sensing images, *i.e.,* without performing simulated degradations. Table. III shows the quantitative comparison of EDiffSR against SOTA methods in terms of NIQE and AG. We can see that EDiffSR achieves the best NIQE performance, illustrating our method can restore natural images that align with human perception in real-world scenarios. In addition, the best AG performance demonstrates that our reconstructed image contains more high-frequency detail information, such as edges and textures.

More intuitively, we display the visual comparison on the NWPU-RESISC45 dataset. The qualitative results are shown in Fig. 10. We can see that the PSNR-driven approach HAT-L exhibits a significantly blurry effect compared to the other approaches. Whereas GAN-based methods such as MSR-GAN produce excessively sharpened results accompanied by pseudo-details. In the diffusion-based methods, SR3 shows limitations in recovering precise edge information of the dense lines on the ground. In contrast, our method demonstrates the clearest preservation of high-frequency texture information, with minimal blurring and artifacts.

TABLE III. Quantitative comparison with state-of-the-art SR models in terms of NIQE, and AG on NWPU-RESISC45 test set. The best performance value is highlighted in **red** while the second best in **blue**.

Metrics	Bicubic	RCAN [3]	HAT-L [4]	MSRGAN [5]	ESRGAN [6]	SPSR [7]	SR3 [8]	IRSDE [9]	EDiffSR
NIQE ↓	20.8032	20.4722	20.3153	17.3114	17.9991	18.0730	17.8176	**16.5357**	**16.3013**
AG ↑	2.3556	3.0296	3.0081	3.3771	3.5170	3.5723	4.0101	**4.2690**	**4.7461**

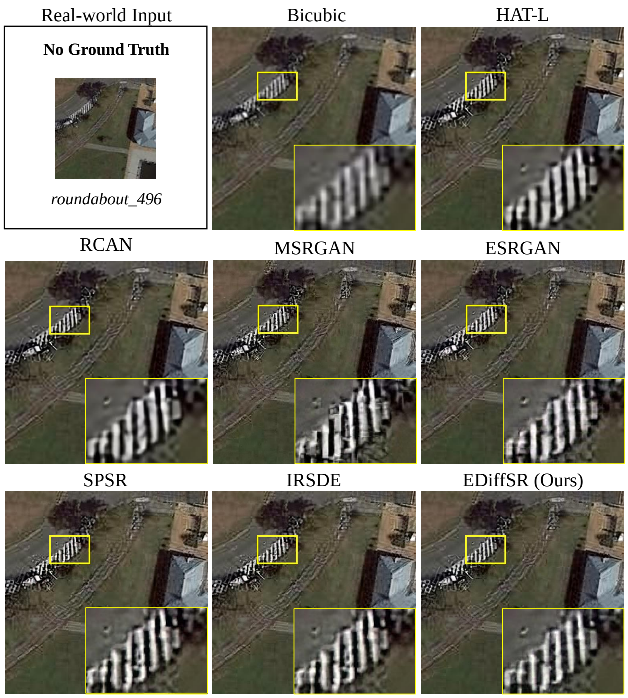

Fig. 10. ×4 visual comparisons with state-of-the-art SR models on NWPU-RESISC45 with real-world degradations. Zoom in for a better view.

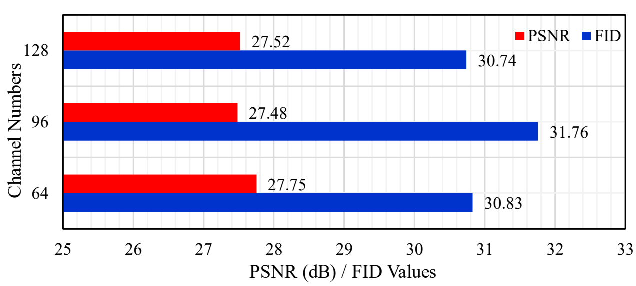

Fig. 11. Ablation analysis of EANet with different channel numbers C.

TABLE IV. Quantitative comparison with state-of-the-art SR models in terms of PSNR, and SSIM across AID, DOTA, and DIOR test sets. The best performance value in each model type is highlighted in **blod** while the second best is in <u>underline</u>.

Method	AID [57]		DOTA [58]		DIOR [59]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [2]	30.65	0.8085	33.64	0.8648	30.63	0.8116
RCAN [3]	**30.82**	<u>0.8121</u>	<u>33.86</u>	<u>0.8680</u>	<u>30.85</u>	<u>0.8159</u>
HAT-L [4]	<u>30.81</u>	**0.8124**	**33.99**	**0.8684**	**30.87**	**0.8161**
MSRGAN [5]	**28.75**	**0.7390**	**29.46**	**0.7825**	**28.84**	**0.7422**
ESRGAN [6]	<u>28.38</u>	<u>0.7272</u>	<u>29.01</u>	<u>0.7716</u>	<u>28.07</u>	<u>0.7086</u>
SPSR [7]	27.71	0.7081	28.00	0.7696	27.46	0.7106
SR3 [8]	26.24	<u>0.6705</u>	27.62	0.6804	26.18	<u>0.6639</u>
IRSDE [9]	<u>27.19</u>	0.6585	<u>28.08</u>	<u>0.7133</u>	<u>27.11</u>	0.6516
EDiffSR (Ours)	**27.40**	**0.6805**	**28.30**	**0.7345**	**27.55**	**0.6823**

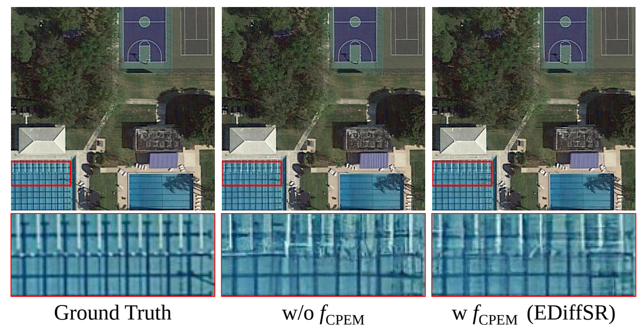

Fig. 12. ×4 visual comparisons of the EDiffSR model without and with the conditional prior enhancement module f_{CPEM} on "img_074" from the DOTA test set. Image restored by the complete EDiffSR shows more high-frequency details than those recovered without f_{CPEM}.

TABLE V. Ablation analysis of EDiffSR with different components. The best FID performance is shown in **blod**.

Methods	f_{CPEM}	EANet	UNet	Param. (M)	FID ↓
Baseline	✗	✗	✓	137.15	32.42
Model-1	✓	✗	✓	137.59	32.68
Model-2	✗	✓	✗	26.34	31.11
EDiffSR (Ours)	✓	✓	✗	26.79	**30.83**

E. Ablation Studies

In this section, we conduct extensive experiments to demonstrate the effectiveness of each component within our EDiffSR. Note that the FID and PSNR values are the average results of the AID, DOTA, and DIOR datasets.

1) Component Analysis of EDiffSR: To investigate the holistic effectiveness of each part within EDiffSR, we remove the conditional prior enhancement (f_{CPEM}), the Efficient Activate Network (EANet) to form the three models reported in Table. V. Note that once we remove the EANet, we replace it with the Vanillia UNet for noise prediction. By comparing Model-1 and EDiffSR, we can find that EANet is superior in improving the FID performance than UNet (30.83 vs. 32.68) while reducing the model size by a large margin (26.31M vs. 137.15M). After adding the f_{CPEM} in Model-2, we observe a slight parameter increase, but the improvement in FID is significant. When the whole f_{CPEM} and EANet are absent in Baseline, the model performs poorly in FID. These results demonstrate that the proposed f_{CPEM} and EANet are able to improve the performance of the diffusion model. Besides, both f_{CPEM} and EANet have low complexity, allowing EDiffSR efficient yet effective.

2) Effectiveness of EANet: We first investigate the impact of varying channel numbers of EANet. As shown in Fig. 11, we find that EDiffSR achieves slightly superior FID performance when $C = 128$ compared to $C = 64$. However, it was observed that EDiffSR yields the highest PSNR results when $C = 64$. To strike a favorable balance between model size and

TABLE VI. Ablation analysis of EANet with different scales of convolution. 3×3, 5×5, and 7×7 represents single-scale design with different DWConv kernels. EDiffSR adopts the multi-scale design and achieves modest improvement.

Methods	3×3	5×5	7×7	EDiffSR (Ours)
FID ↓	30.98	31.17	31.06	**30.83**
PSNR ↑	27.59	27.77	**27.89**	27.75

TABLE VII. Model efficiency analysis with state-of-the-art SR models. The best performance is shown in **blod**.

Methods	Param. (M)	Running Time (s)	FID ↓	PSNR (dB) ↑
EDSR [2]	43.09	0.93	62.50	31.64
RCAN [3]	15.59	0.29	59.99	31.84
HAT-L [4]	40.32	0.76	58.47	**31.89**
MSRGAN [5]	**1.52**	**0.19**	34.88	29.02
ESRGAN [6]	16.70	0.22	33.16	28.49
SPSR [7]	24.79	0.60	35.61	27.72
SR3 [8]	92.56	137.61	43.61	26.68
IRSDE [9]	137.15	27.90	32.42	27.46
EDiffSR (Ours)	26.79	19.26	**30.83**	27.75

performance, we set $C = 64$ in our final EDiffSR. Moreover, we conducted three experiments to assess the impact of multi-scale design in EAB. The results are listed in Table. VI. From the table, we can see that the multi-scale design brings a modest improvement in terms of FID.

3) Effectiveness of Conditional Prior Enhancement: To further illustrate the capability of f_{CPEM} in grasping valuable prior knowledge for accurate SR reconstruction, we provide a visual comparison in Figure 12. From this figure, we can see that the model with f_{CPEM} excels in recovering high-frequency details, such as edges and boundaries. This observation highlights that f_{CPEM} indeed boosts the performance of EDiffSR by exploring an enriched condition with more priors.

4) Model Efficiency: To demonstrate the efficiency of EDiffSR, we conducted a comparison of parameters and inference times, as presented in Table VII. The results indicate that EDiffSR is far more lightweight compared to existing DPM-based SR models. For instance, when compared with IRSDE, EDiffSR achieves an impressive reduction of nearly 80% in model parameters (26.79M vs. 137.15M) while delivering superior performance in both FID (30.83 vs. 32.42) and PSNR (27.75dB vs. 27.46dB). Furthermore, EDiffSR exhibits faster inference compared to existing DPM-based models. When compared to SR3, EDiffSR is 7 times faster (19.26s vs. 137.61s) in the diffusion sampling process, making it more practical for real-world applications.

VI. Conclusion

In this paper, we devise an efficient diffusion probabilistic model (EDiffSR) to generate perceptual-pleasant SR results of remote sensing images. The proposed efficient activation network (EANet) shows superior performance against vanilla UNet in noise prediction and is more lightweight. In particular, rather than employing the interpolated condition, a condition prior enhancement module is designed to explore the potential priors from LR input, which significantly boosts the reconstruction performance. Rigorous quantitative and qualitative evaluation on AID, DOTA, DIOR, and NWPU-RESISC45 datasets demonstrated our EDiffSR outperforms state-of-the-art CNN-based, GAN-based, and Diffusion-based SR methods in data distribution and perceptual quality.

Nevertheless, our EDiffSR does exhibit some shortcomings. Firstly, the sampling process of the diffusion model consumes massive computational costs, which hinders its real-time application. Secondly, EDiffSR does not consider the multiple degradations involved in remote sensing images, resulting in limited adaptability to real-world scenes. Therefore, more efforts should be paid to speed up the sampling process of the diffusion model in the future direction. Moreover, we consider extending our EDiffSR to blind SR issues, thus improving its generalization in real-world scenarios.

References

[1] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[3] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[4] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 367–22 377.

[5] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[6] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 1–17.

[7] C. Ma, Y. Rao, J. Lu, and J. Zhou, "Structure-preserving image super-resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7898–7911, 2021.

[8] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023.

[9] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations," in *International Conference on Machine Learning*. PMLR, 2023.

[10] J. He, Q. Yuan, J. Li, Y. Xiao, D. Liu, H. Shen, and L. Zhang, "Spectral super-resolution meets deep learning: Achievements and challenges," *Information Fusion*, p. 101812, 2023.

[11] Y. Xiao, Q. Yuan, Q. Zhang, and L. Zhang, "Deep blind super-resolution for satellite video," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[12] D. Liu, J. Li, Q. Yuan, L. Zheng, J. He, S. Zhao, and Y. Xiao, "An efficient unfolding network with disentangled spatial-spectral representation for hyperspectral image super-resolution," *Information Fusion*, vol. 94, pp. 92–111, 2023.

[13] R. Dian, A. Guo, and S. Li, "Zero-shot hyperspectral sharpening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 650–12 666, 2023.

[14] D. He and Y. Zhong, "Deep hierarchical pyramid network with high-frequency-aware differential architecture for super-resolution mapping," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[15] D. He, Q. Shi, X. Liu, Y. Zhong, G. Xia, and L. Zhang, "Generating annual high resolution land cover products for 28 metropolises in china based on a deep super-resolution mapping network using landsat imagery," *GIScience & Remote Sensing*, vol. 59, no. 1, pp. 2036–2067, 2022.

[16] J. He, Q. Yuan, J. Li, Y. Xiao, and L. Zhang, "A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, pp. 131–144, 2023.

[17] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by cnn denoiser," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 3, pp. 1124–1135, 2020.

[18] Y. Xiao, Y. Wang, Q. Yuan, J. He, and L. Zhang, "Generating a long-term (2003- 2020) hourly 0.25° global pm2. 5 dataset via spatiotemporal downscaling of cams with deep learning (deepcams)," *Science of The Total Environment*, vol. 848, p. 157747, 2022.

[19] Y. Xiao, Q. Yuan, J. He, Q. Zhang, J. Sun, X. Su, J. Wu, and L. Zhang, "Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer," *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, p. 102731, 2022.

[20] Q. Zhang, Y. Zheng, Q. Yuan, M. Song, H. Yu, and Y. Xiao, "Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2023.

[21] Y.-C. Miao, X.-L. Zhao, X. Fu, J.-L. Wang, and Y.-B. Zheng, "Hyperspectral denoising using unsupervised disentangled spatiospectral deep priors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[22] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Hierarchical dense recursive network for image super-resolution," *Pattern Recognition*, vol. 107, p. 107475, 2020.

[23] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[24] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[25] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3517–3526.

[26] J. He, Q. Yuan, J. Li, Y. Xiao, X. Liu, and Y. Zou, "Dster: A dense spectral transformer for remote sensing spectral super-resolution," *International Journal of Applied Earth Observation and Geoinformation*, vol. 109, p. 102773, 2022.

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[29] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[30] H. Wu, N. Ni, S. Wang, and L. Zhang, "Conditional stochastic normalizing flows for blind super-resolution of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[31] Y. Miao, L. Zhang, L. Zhang, and D. Tao, "Dds2m: Self-supervised denoising diffusion spatio-spectral model for hyperspectral image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 086–12 096.

[32] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[33] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4491–4500.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[35] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," *arXiv preprint arXiv:2303.09472*, 2023.

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

[37] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[38] 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," *Information Fusion*, vol. 96, pp. 297–311, 2023.

[39] S. Zhang, Q. Yuan, J. Li, J. Sun, and X. Zhang, "Scene-adaptive remote sensing image super-resolution using a multiscale attention network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4764–4779, 2020.

[40] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[41] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[42] S. Chen, L. Zhang, and L. Zhang, "Msdformer: Multiscale deformable transformer for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[43] D. Zhang, J. Shao, X. Li, and H. T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5183–5196, 2021.

[44] Y. Xiao, Q. Yuan, K. Jiang, X. Jin, J. He, L. Zhang, and C.-w. Lin, "Local-global temporal difference learning for satellite video super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–16, 2023, doi: 10.1109/TCSVT.2023.3312321.

[45] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[46] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 31, pp. 7252–7263, 2022.

[47] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Rgb-induced feature modulation network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[48] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3633–3643, 2019.

[49] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced gan for remote sensing image superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5799–5812, 2019.

[50] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6792–6810, 2018.

[51] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 5662–5673, 2022.

[52] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, and B. Lu, "Diffusion model with detail complement for super-resolution of remote sensing," *Remote Sensing*, vol. 14, no. 19, p. 4834, 2022.

[53] L. Han, Y. Zhao, H. Lv, Y. Zhang, H. Liu, G. Bi, and Q. Han, "Enhancing remote sensing image super-resolution with efficient hybrid conditional diffusion model," *Remote Sensing*, vol. 15, no. 13, p. 3452, 2023.

[54] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[55] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[56] P. E. Kloeden, E. Platen, P. E. Kloeden, and E. Platen, *Stochastic differential equations*. Springer, 1992.

[57] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[58] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[59] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS*

journal of photogrammetry and remote sensing, vol. 159, pp. 296–307, 2020.

[60] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[61] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[62] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[64] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[65] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

Yi Xiao received the B.S. degree from the School of Mathematics and Physics, China University of Geosciences, Wuhan, China, in 2020. He is pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan.

His major research interests are remote sensing image super-resolution and computer vision. More details can be found at https://xy-boy.github.io/

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is a Professor. He has published more than 90 research papers, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTION ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was a recipient of the Youth Talent Support Program of China in 2019, the Top-Ten Academic Star of Wuhan University in 2011, and the recognition of Best Reviewers of the IEEE GRSL in 2019. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is an associate editor of 5 international journals and has frequently served as a referee for more than 40 international journals for remote sensing and image processing.

Kui Jiang (Member, IEEE) received the Ph.D. degree in the school of Computer Science, Wuhan University, Wuhan, China, in 2022. He is currently an associate professor with the school of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. He received the 2022 ACM Wuhan Doctoral Dissertation Award. His research interests include image/video processing and computer vision.

Jiang He received the B.S. degree in remote sensing science and technology from faculty of geosciences and environmental engineering in Southwest Jiaotong University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree in School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include hyperspectral super-resolution, image fusion, quality improvement, remote sensing image processing and deep learning.

Xianyu Jin received the B.S. degree in geodesy and geomatics engineering from Wuhan University, Wuhan, China, in 2019, where he is pursuing the M.S. degree with the School of Geodesy and Geomatics.

His research interests include video super-resolution, deep learning, and computer vision.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently a "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China at the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He was a Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He has published more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) Highly Cited Author. He holds 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE GRSS 2014 Data Fusion Contest. His students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is also the Founding Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter. He also serves as an associate editor or an editor for more than ten international journals. He is also serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.