WaveDiffUR: A Wavelet-Domain Diffusion Model for Ultra-Resolution in Remote Sensing

Yue Shi, Liangxiu Han*, Lianghao Han, Darren Dancey, Xueqin Zhang

***Abstract*—Deep learning (DL) has significantly advanced super-resolution (SR), a technique that enhances low-quality images by reconstructing fine details. However, most DL-based SR methods struggle at high magnification levels (e.g., $\times 4$ or higher) due to dramatically increased ill-posedness. To overcome this, we define high-magnification SR as an ultra-resolution (UR) problem and introduce WaveDiffUR, a novel wavelet-domain diffusion model designed for extreme-scale image reconstruction. WaveDiffUR decomposes the UR process into sequential steps, first restoring low-frequency wavelet details for global consistency and then refining high-frequency components for sharper textures. By integrating pre-trained SR models as modular components, it reduces ill-posedness and ensures adaptability across different applications. Unlike existing SR approaches, which struggle with fixed boundary conditions at extreme magnifications, WaveDiffUR incorporates the cross-scale pyramid (CSP) constraint, an adaptive framework that dynamically refines low- and high-frequency wavelet details to maintain consistency and high fidelity. Extensive experiments demonstrate that WaveDiffUR with CSP notably enhances spatial accuracy and consistently generates high-frequency details with remarkable fidelity during the SR process. Evaluations are conducted across two benchmark evaluation datasets and four additional independent datasets. The empirical results reveal that, as magnification scales from $\times 8$ to $\times 128$, WaveDiffUR achieves an average degradation rate in PSNR, NIQE, and SRE of only 19.1%—the best performance among all benchmarked models—while consistently delivering sharper images characterized by superior spatial fidelity. By enabling scalable, high-fidelity ultra-resolution, WaveDiffUR opens new possibilities for remote sensing applications, including environmental monitoring, urban planning, disaster response, and precision agriculture.**

***Index Terms*—Remote sensing Image Super Resolution; Ultra Resolution (UR); Diffusion Model; Wavelet Transformation; Stochastic Differential Equation (SDE); Multi-Scale Generative AI**

I. Introduction

REMOTE sensing image super-resolution (SR) remains a persistent challenge and continues to be a vibrant research topic in both computer vision [1] and geosciences [2]. SR aims to reconstruct high-resolution (HR) remote sensing images with realistic spectral-spatial details from low-resolution (LR) data [3], typically acquired from aerial platforms (1 - 10 m resolution) [4] or space-based platforms ($\geq$ 10m resolution) [5]. However, most existing SR research focuses on fixed and low-magnification scales (e.g., $\times 2$ or $\times 4$) [6]–[8], which fails to meet the growing demand for high-magnification SR in many Earth observation tasks. In this study, we define high-magnification SR as an ultra-resolution (UR) problem. The high magnification UR introduces big challenges, as small reconstruction errors can lead to significant detail loss or artifacts over large areas. For example, land-cover mapping typically requires a spatial resolution of 1–2 meters, necessitating $\times 8$ UR for 10-meter Sentinel-2 data or $\times 16$ UR for 30-meter Landsat-8 data, where the sub-pixel reconstruction error around boundaries can shift class edges, causing narrow features such as roads, hedgerows, or riparian strips to disappear and leading to systematic misclassification of landcover classes [9]. Similarly, precision agriculture demands resolutions higher than 1 meter, translating to $\times 16$ to $\times 32$ UR for Satellite data, where spectral–spatial biases introduced by UR reconstruction can distort canopy patterns and propagate errors into vegetation indices, resulting in inaccurate crop stress detection [10].

Unlike natural image super-resolution, remote sensing SR presents unique challenges due to complex spatial heterogeneity and the presence of mixed pixels. These factors further exacerbate the ill-posed nature of high-magnification SR, where a given low-resolution (LR) input can correspond to

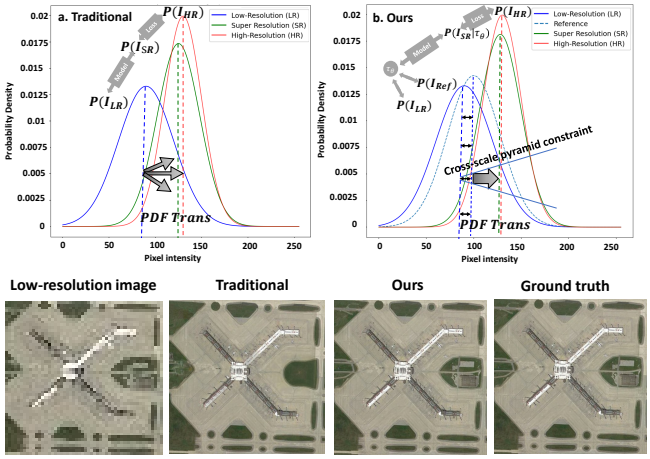

Fig. 1. Comparison of probability density function (PDF) transitions in (a) a traditional diffusion-based SR model [7] and (b) the proposed WaveDiffUR model. Results at $\times 16$ UR, shown as a representative example, illustrate that the proposed method alleviates the ill-posedness of remote sensing UR tasks compared to traditional approaches.

This work was supported in part by the Biotechnology and Biological Sciences Research Council (BBSRC) under Grant BB/Y513763/1, and in part by Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/X013707/1, (Corresponding author: Liangxiu Han.).

Yue Shi, Liangxiu Han, Darren Dancey are with the Department of Computing and Mathematics, Manchester Metropolitan University, M1 5GD Manchester, U.K. (e-mail: y.shi@mmu.ac.uk; l.han@mmu.ac.uk; d.dancy@mmu.ac.uk).

Lianghao Han is with the Department of Computer Science, Brunel University, London UB8 3PH, U.K.

Xueqin Zhang, School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China.

an infinite number of possible high-resolution (HR) solutions [11].

Most deep learning-based SR methods attempt to mitigate this issue by training neural networks to model probability density function (PDF) transitions in the pixel-wise representation space, effectively mapping the PDF of LR images to their HR counterparts. However, these methods remain limited in handling extreme magnification scales, where the solution space grows exponentially. Figure 1 presents a comparative analysis of popular PDF-based deep learning SR models, including our previous approach [12], highlighting their performance differences at increasing magnification levels.

Among existing approaches, generative adversarial networks (GANs) use adversarial learning between a generator and discriminator to synthesize SR images with realistic high-resolution (HR) details that are missing in low-resolution (LR) inputs. This process aims to align the probability density function (PDF) of SR images with that of the HR counterparts [13], [14]. Variational autoencoders (VAEs) take a different approach, encoding the LR PDF into a latent space, then generating SR images via sampling, ensuring the reconstructed SR PDF aligns with the HR image distribution [15].

Despite their success, these methods struggle at high magnification levels due to the inherently ill-posed nature of SR. Most SR models are trained on LR-HR image pairs with low magnification rates (e.g., $\times 2$ or $\times 4$), which provide effective cross-scale representations for learning-based models. Our previous study [12] explored a GAN-based approach for high-magnification SR and found that once the magnification exceeds $\times 8$, SR quality deteriorates due to mode collapse and perceptual artifacts. This degradation arises from the adversarial nature of GANs, which are notoriously difficult to converge at extreme scales due to the increased complexity of PDF transitions [13]. Additionally, at high magnification levels, the discontinuity in cross-scale representations reduces the model's ability to estimate HR outputs from LR inputs, making the LR-to-HR PDF transition significantly more complex and unpredictable.

Recently, diffusion models (DMs) [16] have gained attention in image restoration and have demonstrated promising results in remote sensing SR [7], [8], [17]. The strength of DMs lies in their denoising diffusion process, which gradually refines the LR PDF into the HR PDF through small, incremental noise removal steps. Unlike GANs, DMs provide a well-defined probabilistic framework, avoiding training instability and mode collapse. However, DM-based UR is far more ill-posed than ordinary SR [18], because the extreme magnification factors greatly expand the range of plausible high-frequency details; consequently, diffusion-based methods—whose sampling process is inherently stochastic—often struggle to maintain coherent spectral-spatial information across large homogeneous regions [19].

To address this challenge, we introduce the cross-scale pyramid (CSP) boundary condition, which captures spectral-spatial unmixing rules across different magnification levels. Building on this concept, we formulate the UR process as a conditional diffusion stochastic differential equation (SDE). This framework enables low-frequency fidelity enhancement by reconstructing global details while ensuring high-frequency consistency refinement by restoring local textures. To solve this SDE, we propose WaveDiffUR, a wavelet-based diffusion UR solver that operates in the wavelet domain, allowing it to mitigate the ill-posed nature of UR. The WaveDiffUR framework, shown in Figure 2, seamlessly integrates pre-trained SR pipelines as plug-and-play modules to generate cross-scale conditions, reducing the computational cost of training new models from scratch. However, using fixed boundary conditions throughout the UR process can limit constraint capacity, degrading the consistency and fidelity of the results. To overcome this limitation, we introduce a dynamically updated cross-scale condition named CSP. This serves as a variable boundary condition for the SDE solver, continuously compressing information from adjacent UR sub-processes. By doing so, CSP guides WaveDiffUR to produce accurate UR results with realistic spectral-spatial consistency. Experimental results demonstrate that the baseline WaveDiffUR model without CSP exhibits high performance in terms of usability, adaptability, and cost-effectiveness. Moreover, the enhanced CSP-WaveDiffUR model effectively captures the unmixing rules of realistic spectral-spatial details, thereby improving UR efficiency and robustness in handling high-magnification SR tasks.

The primary contributions of this work are as follows:

(i) We pioneer a solution to the Ultra-Resolution (UR) problem, termed WaveDiffUR, which decomposes the complex UR process into finite sub-processes. It leverages pre-trained SR models to their fullest potential, enhancing the usability, adaptability, and cost-effectiveness of the UR process. To the best of our knowledge, this is the first work that explicitly addresses the complex and ill-posed UR problem.

(ii) We propose an improved version of the SDE solver, CSP-WaveDiffUR, to address the degradation issue caused by fixed boundary conditions in diffusion-based UR SDEs. This model dynamically updates boundary conditions during each UR sub-process, ensuring more stable and high-fidelity UR results.

By addressing fundamental challenges in ultra-resolution, WaveDiffUR opens new opportunities for practical remote sensing applications, including environmental monitoring, urban planning, disaster response, and precision agriculture. This study presents a scalable, cost-effective, and high-fidelity approach to advancing remote sensing capabilities at unprecedented magnification levels.

The remainder of this paper is organized as follows: Section II reviews the related work on diffusion models and remote sensing image super-resolution. The methodology is detailed in Section III, including the main framework of the proposed WaveDiffUR method. Section IV presents the experimental results. Finally, Section V concludes this work and highlights future directions.

II. Related Work

A. Remote Sensing Image Super-Resolution

To transform low-resolution remote sensing images into high-resolution counterparts, considerable efforts have been

made to improve image fidelity and restore details. Traditional methods primarily rely on fusion techniques, such as wavelet transform and spectral mixing analysis, where the spatial information from high-resolution images is leveraged to enhance the spatial details of low-resolution images effectively. For instance, Zhang *et al.* [20] introduced a remote sensing image fusion technique based on the 3D Wavelet Transform (3DWT), where 3DWT effectively harnesses spectral information to generate high-quality fused spectral-spatial details.

In recent years, significant advancements have been made in deep learning-based approaches to the SR problem. For example, Zheng *et al.* [21] applied a spectral-spatial attention mechanism to neural networks for panchromatic sharpening of remote sensing images, enabling the networks to adaptively learn both spatial and spectral details. Li *et al.* [22] designed a spectral super-resolution framework by learning a cross-scale relationship and achieved a satisfactory result. More recently, they transferred the spectral unmixing into the super-resolution and hence proposed an effective coupled unmixing framework [23].

Deep learning-based approaches for improving the spatial resolution of remote sensing images have primarily followed two strategies: fusion with high-spatial-resolution images [24]–[26] and single-image-based SR [27], [28]. Fusion-based SR techniques leverage external prior information to reconstruct images with finer textures. In contrast, single-image-based SR techniques operate without auxiliary data, offering greater practical feasibility. For instance, Mei *et al.* [29] developed a 3D Fully Convolutional Neural Network (3D-FCNN) for drone image super-resolution, incorporating an upsampling process in the early stages. Jiang *et al.* [30] proposed the Single Sub-Image Progressive Super-Resolution (SSPSR) model, which employs a progressive sampling approach: first for grouped sub-images, followed by the fusion of interpolated sub-images to construct the entire image. This approach enhances feature extraction in HSIs and improves overall training stability. However, it introduces additional requirements, such as more precise modeling and intricate network design at each stage.

The Generative Adversarial Network (GAN) is another deep learning-based model that has gained significant attention in the field of super-resolution. GANs are particularly valued for their ability to model complex data distributions, enabling the generation of high-resolution (HR) images that closely resemble real-world data in both quality and perceptual characteristics. When incorporated into the SR process, GANs generate HR images with enhanced perceptual quality. Xiong *et al.* [31] proposed an improved Super-Resolution Generative Adversarial Network (SRGAN) featuring a revised loss function and an optimized network architecture. These modifications enhance training stability and improve generalization performance. Shi *et al.* [12] proposed the Latent Encoder-Integrated GAN (LE-GAN), which incorporates self-attention mechanisms to enhance feature extraction in the generator and stabilize the training process.

In parallel, diffusion probabilistic models (DPMs) have emerged as another promising approach for super-resolution tasks. DPMs generate high-quality data distributions through a structured and well-defined probabilistic diffusion process, mitigating the training instability often seen in GANs. Recently, Saharia *et al.* [32] proposed a DPM-based super-resolution method, using a UNet architecture as the denoiser to iteratively refine image generation. Luo *et al.* [33] further enhanced diffusion-based SR by introducing stochastic differential equations (SDEs) to more accurately model the degradation process in diffusion. These advancements highlight the potential of diffusion models in addressing complex SR challenges. A detailed investigation of diffusion-based super-resolution is presented in the following section.

B. Diffusion-based Image Super-Resolution

Diffusion-based models, particularly denoising diffusion probabilistic models (DDPMs) [16], have shown that iterative denoising can yield high-quality image restoration results, including super-resolution [7], [34], [35], inpainting [36], [37], and deblurring [38], [39]. For example, Kawar *et al.* [40] introduced Denoising Diffusion Restoration Models (DDRM), which utilize a pre-trained diffusion model to solve various linear inverse problems, demonstrating superior performance across multiple image restoration tasks. Wang *et al.* [41] proposed DR2, a Diffusion-Based Robust Degradation Remover for Blind Face Restoration. DR2 first employs a pre-trained diffusion model for coarse degradation removal, followed by an enhancement module designed for finer blind face restoration. Guo *et al.* [42] developed ShadowDiffusion, which employs an unrolled diffusion model to tackle the challenging task of shadow removal by progressively refining results using degradation and generative priors.

The basic principle of diffusion-based image super-resolution involves the use of a Markov chain to model the transformation of high-resolution (HR) image data into noise and back again [40]. It consists of two opposing processes: the forward process (diffusion process) and the reverse process (denoising with a condition). The forward process gradually corrupts an HR image through a Markov chain, transforming the HR image distribution into a stochastic Gaussian noise distribution by progressively adding noise. This process effectively creates a dataset of noisy images representing the HR data in a stochastic space. In the reverse process, the HR image is reconstructed from the noisy data using the corresponding low-resolution image as a conditioning factor to systematically guide noise removal. The model progressively refines the noisy data, transforming it back into the HR distribution. Through this denoising process, the conditional diffusion model ensures spatial and spectral consistency with the LR input while reconstructing fine details from the noisy data.

The forward noising process in diffusion-based image super-resolution is governed by the following stochastic differential equation (SDE) [43]:

$$dx = \bar{f}(x,t)dt + \bar{g}(t)dw, \tag{1}$$

where $\bar{f}(x,t)$ is the linear drift function governing the rate at which the HR image data x is perturbed at time step t, $\bar{g}(t)$ is the scalar diffusion coefficient associated with t, and w denotes the standard Wiener process.

Using Anderson's theorem [43], [44], the reverse diffusion process in SR can be expressed as a reverse stochastic differential equation (SDE):

$$dx = \left[\bar{f}(x,t) - \bar{g}(t)^2 \nabla_x \log p_t(x \mid y)\right] dt + \bar{g}(t) d\bar{w}, \quad (2)$$

where x represents the reconstructed HR image, y is a conditioning variable typically derived from the LR image or intermediate features, dt denotes the infinitesimal negative time step, and $\bar{w}$ represents the reverse Wiener process. The reverse SDE governs the generative process through the score function $\nabla_x \log p_t(x \mid y)$ and minimizes the following denoising score-matching objective:

$$\min_{\theta} \mathbb{E}_{t\sim U(\epsilon,1),\, x_0\sim p_0(x|y),\, x_t\sim p_{0,t}(x_t|x_0)} \left\| s_\theta(x_t,t) - \nabla_{x_t} \log p_{0,t}(x_t \mid x_0) \right\|_2^2, \quad (3)$$

Once the parameter θ for the score function is estimated, the score function $\nabla_x \log p_t(x \mid y)$ in Eq. 2 can be replaced with $s_\theta(x_t, t)$ to solve the reverse SDE.

C. Diffusion Process in the Wavelet Domain

The wavelet transform is highly effective in reducing spatial dimensions while retaining critical information, distinguishing it from transformation techniques such as the Fast Fourier Transform (FFT) and the Discrete Cosine Transform (DCT), which may suffer from information loss during transformation. This advantage has led to its adoption in diffusion-based super-resolution (SR). Studies such as that by Jiang et al. [45] highlight the benefits of performing diffusion operations in the wavelet domain rather than directly in the image space. This approach enhances content reconstruction and reduces the disparity between ground-truth HR and SR domains. By integrating wavelet transforms with diffusion-based models, these methods achieve superior super-resolution performance and improved evaluation metrics.

In 2D applications, a low-resolution image $x \in I_{LR}$ can be decomposed into four sub-bands using the 2D discrete wavelet transform (2D-DWT) with Haar wavelet functions, as follows:

$$\{A_{LR}, V_{LR}, H_{LR}, D_{LR}\} = \text{2D-DWT}(I_{LR}), \quad (4)$$

where $A_{LR} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$ represents the low-frequency component of the input image, while $V_{LR}, H_{LR}, D_{LR} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$ correspond to the vertical, horizontal, and diagonal high-frequency components of the input image, respectively. Specifically, the low-frequency coefficient A_{LR} preserves the global structural information of the original image, acting as a downsampled representation, while the high-frequency coefficients V_{LR}, H_{LR}, and D_{LR} capture sparse local details that contribute to the fine details and texture of the image.

In image super-resolution, reconstructing images by upscaling high-frequency coefficients helps restore the fine details of the original images. In contrast, the low-frequency coefficient primarily governs the global structure, ensuring greater consistency with the original images. In wavelet-domain super-resolution, particularly at ultra-resolution scales, accurately upscaling the low-frequency (LL) sub-band to match the high-resolution scene establishes the global geometry and structure consistency, furnishing the scaffold on which the high-frequency sub-bands can later be synthesized to restore fine textures and edges [46].

III. Methodology

We hypothesize that the ill-posed UR process can be modeled as a stochastic diffusion process for spectral-spatial unmixing, systematically addressing the interplay between spectral fidelity and spatial consistency. Our primary goal is to address the ill-posed UR problem using a diffusion UR SDE and generate high-fidelity, spatially consistent UR images from their low-resolution counterparts. The proposed UR process is extended to a finite number of SR steps by formulating them as the solution to a conditional diffusion SDE, constrained by a pyramid-shaped multi-scale spectral-spatial unmixing rule.

Inspired by Jiang *et al.* [45], we adopt a wavelet-domain implementation for the proposed UR process to enhance computational efficiency. Wavelet decomposition inherently provides multi-scale components, aligning with our UR stepwise restoration strategy. In addition, Wavelet decomposition divides the UR mapping into multiple frequency bands, transforming a single highly ill-posed problem into a set of smaller, more manageable sub-problems. Low-frequency sub-bands capture large-scale structures that are easier to infer from low-resolution input, while high-frequency sub-bands represent fine details, which are reconstructed conditionally. This separation regularises the solution space by enforcing that each band's reconstruction must be physically consistent when combined. The workflow of the proposed UR approach is illustrated in Figure 2. The following sections provide a detailed explanation of the proposed method.

A. Wavelet-Domain Diffusion UR (WaveDiffUR) SDE Solver

According to our hypothesize, the UR process in the wavelet domain operates as a stochastic diffusion process for spectral-spatial unmixing, transitioning from low-resolution to high-resolution space. This is achieved through a series of SR steps formulated as solutions to conditional diffusion SDEs.

Our proposed WaveDiffUR solver addresses the conditional diffusion UR SDEs to enhance both low-frequency and high-frequency details. The solver ensures that spectral-spatial consistency mitigates the ill-posed nature of the UR task by dynamically adapting to multi-scale constraints across magnification levels, and preserves the integrity of fine details during high-magnification upscaling.

Each SR step iteratively refines the wavelet components, progressively generating UR wavelet representations for each frequency band. Low-frequency components benefit from pre-trained SR modules to enhance global structures, while high-frequency components are upscaled to improve texture and detail. After processing the wavelet components, the Inverse Discrete Wavelet Transform (IDWT) reconstructs the image in the spatial domain. This final reconstruction step ensures that the synthesized HR image maintains the desired fidelity and consistency with HR ground-truth data.

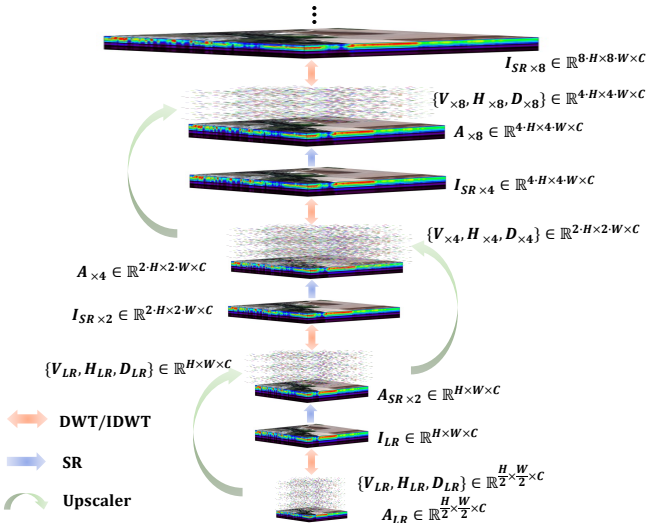

Fig. 2. An illustration of the proposed self-cascade UR pyramid framework, consisting of: 1) DWT/IDWT: cyclically decomposes a low-resolution image into the wavelet domain and restores the high-resolution image from the upscaled wavelet-domain components. 2) SR pipeline: integrates a plug-and-play tunable SR module into the framework to reconstruct the low-frequency wavelet components of a high-resolution image from its low-resolution counterpart. 3) Upscaler: progressively adjusts the high-frequency wavelet components of low-resolution images to align with those of higher-resolution images.

The low-frequency and high-frequency components after DWT in WaveDiffUR can be computed as follows:

Low-Frequency Components. The low-frequency components of the target image I_{SR} can be expressed as a function of the low-resolution image I_{LR} and the upscaling process. Mathematically, the ideal low-frequency components A_{SR} of the SR image can be expressed as:

$$A_{SR} \sim p(A_{SR}|TRANS(I_{LR})) + \epsilon, \tag{5}$$

where $TRANS$ is the probability density transformation function that maps I_{LR} to the domain of A_{SR}, and independent Gaussian noise ϵ accounts for uncertainties and imperfections in the transformation process. The output of $TRANS(I_{LR})$ serves as a conditioning variable that constrains the probability flow of A_{SR} during the UR process. The low-frequency components follow the conditional probability distribution $p(A_{SR}|TRANS(I_{LR}))$.

However, due to the ill-posed nature of the transformation function, obtaining an analytical form of $TRANS$ in Eq. 5 is challenging. Consequently, we approximate the transformation function using a neural network τ_θ.

$$\tau_\theta(I_{LR}) = SRMapping(p(I_{LR})) + \epsilon \rightarrow \mathbb{R}^{A_{SR}}, \tag{6}$$

where $SRMapping(p(I_{LR}))$ represents any arbitrary pre-trained SR pipeline that maps I_{LR} to A_{SR}, expressed as $\mathbb{R}^{I_{LR}} \rightarrow \mathbb{R}^{A_{SR}}$. Mathematically, $\tau_\theta(I_{LR})$ serves as a projection-based conditioning mechanism to mitigate the ill-posed nature of the problem. Thus, A_{SR} is sampled from the conditional probability distribution $p(A_{SR}|\tau_\theta)$. The low-frequency SR process is further formulated as a reverse diffusion SDE:

$$dA_{SR} = \left[\bar{f}(A_{SR,t}, t) - \bar{g}(t)^2 \nabla_{A_{SR,t}} \log p_t(A_{SR,t} \mid \tau_\theta(I_{LR}))\right] dt + \bar{g}(t) d\bar{w}, \tag{7}$$

where $\bar{f}(\cdot)$ defines the linear drift function, $\bar{g}(\cdot)$ presents a scalar diffusion coefficient, and $\bar{w}$ denotes the standard Wiener process.

The conditional score function, $\nabla_{I_{SR}} \log p(I_{SR}|\tau_\theta(I_{LR}))$, necessitates model retraining whenever the cross-scale condition $\tau_\theta(I_{LR})$ is updated. This dependency restricts the generalization capability of the low-frequency SR process.

To overcome this limitation, inspired by [44], we adopt an unconditional score function, $\nabla_x \log p_t(A_{SR})$, which is independent of $\tau_\theta(I_{LR})$. Instead, $\tau_\theta(I_{LR})$ is introduced as an auxiliary input in the score-based framework, guiding the denoising process while preserving the model's adaptability. More specifically, it can be described as follows:

$$A'_{SR,t-1} = \bar{f}(A_{SR,t}, s_\theta) + \bar{g}(A_{SR,t}) \cdot z, \quad z \sim N(0, I), \tag{8}$$

$$A_{SR,t-1} = \alpha \cdot A'_{SR,t-1} + b_t, \tag{9}$$

where α and b_i are functions of $\tau_\theta(I_{LR})$ and $A_{SR,t}$. Note that Eq. 8 corresponds to the unconditional reverse diffusion term in Eq. 2, whereas Eq. 9 incorporates the projection-based condition.

High-Frequency Wavelet Components. In WaveDiffUR, the upscaler model is responsible for transforming the low-resolution wavelet components $VHD_{LR} = \{V_{LR}, H_{LR}, D_{LR}\}$ into their high-resolution counterparts $VHD_{SR} = \{V_{SR}, H_{SR}, D_{SR}\}$. This process is formulated as an upscaling transformation followed by the addition of independent Gaussian noise. The upscaler generates VHD_{SR} as:

$$VHD_{SR} \sim \mathcal{N}\left(\mathbb{U}(VHD_{LR}), sd^2 \mathbb{U}\mathbb{U}^\top + \sigma^2 I\right), \tag{10}$$

where $\mathbb{U}$ represents the upscaling function, sd denotes the standard deviation associated with VHD_{LR}, and $\sigma^2 I$ corresponds to independent Gaussian noise. The upscaling function $\mathbb{U}$ can be approximated using a pre-trained SR model. Thus, Eq. 10 can be rewritten as:

$$VHD_{SR} \sim \mathcal{N}\left(\mathbb{U}_{SR}(VHD_{LR}), sd^2 \mathbb{U}_{SR}\mathbb{U}_{SR}^\top + \sigma^2 I\right), \tag{11}$$

The WaveDiffUR method is detailed in Algorithm 1.

B. Cross-Scale Pyramid (CSP) Constraint WaveDiffUR Solver

In the WaveDiffUR SDE, re-using a fixed boundary condition $\tau_\theta(I_{LR})$ leads to cumulative errors because the underlying UR problem is ill-posed. To stabilise the process, we replace that static boundary with a dynamically updated Cross-Scale Pyramid (CSP) constraint. This CSP constraint is conceptually inspired from the pyramid spectral-spatial unmixing

Algorithm 1 WaveDiffUR SDE solver

Input: Low-resolution image I_{LR}, pre-trained SR pipeline SR, pre-trained U-Net s_θ, current resolution r, target UR resolution R, SR rate k, linear drift function $\bar{f}(\cdot)$, scalar diffusion coefficient $\bar{g}(\cdot)$.
Output: High-resolution image I_{UR}.
Compute UR rate $K = R/r$ and number of UR steps $d = K/k$.
Apply Discrete Wavelet Transform (DWT).
$\{A_{LR}, VHD_{LR}\} = DWT(I_{LR})$
$\tau_\theta(I_{LR}) = SRMapping(p(I_{LR})) + \epsilon$
for i = 1 to d (UR steps) **do**
  for $t = T$ to 1 (reverse diffusion process) **do**
    Perform low-frequency wavelet super-resolution (SR)
    $A'_{SR,t-1} = \bar{f}(A_{SR,t}, s_\theta) + \bar{g}(A_{SR,t}) \cdot z$
    $A_{SR,t-1} = \alpha \cdot A'_{SR,t-1} + b_i$
  end for
  Perform high-frequency wavelet restoration
  $VHD_{SR} \sim \mathcal{N}\left(\mathbb{U}_{SR}(VHD_{LR}), sd^2\mathbb{U}_{SR}\mathbb{U}_{SR}^\top + \sigma^2 I\right)$
  $I_{SR} = IDWT(\{A_{SR,0}, VHD_{SR}\})$
  if $i \neq d$ **then** ▷ Update for next iteration
    $I_{LR} = I_{SR}$
  end if
  if $i = d$ **then** ▷ Final assignment
    $I_{UR} = I_{SR}$
  end if
end for

rule across scales [47]. By doing such pyramid-shaped iterations, it preserves the spectral proportions (avoiding spectral distortion) while increasing spatial detail. In other words, CSP guides the UR process to 'unmix' the pixel information consistently as resolution increases. To initialize the unmix rule, the CSP introduces a reference image $I_{ref} \in \mathbb{R}^{H' \times W' \times c}$ $(H' \geq H,\ W' \geq W)$ and models the joint statistics of the low-resolution input and its reference:

$$\tau_\theta(I_{LR}, I_{ref}) = T_\theta \cdot p(I_{ref}, I_{LR}) + \epsilon \rightarrow \mathbb{R}^{A_{SR}}, \tag{12}$$

where T_θ is a learnable transform, and ϵ is independent Gaussian noise accounting for uncertainties. We denote this low-frequency constraint by τ_θ^{lf}. Because θ is updated at every magnification level, τ_θ^{lf} always stays close to the target domain, providing a stronger guide for the inverse diffusion step:

$$A_{SR,t} \sim p(A_{SR,t}|\tau_\theta^{lf}) + \epsilon, \tag{13}$$

The CSP-WaveDiffUR solver proceeds scale by scale, using two steps: 1) Low-frequency pass and 2) High-frequency constraints. For low-frequency pass, we use Eqs. 7–9 with the current τ_θ^{lf} to recover the low-frequency wavelet band A_{SR}.

For high-frequency constraints, the high-frequency subbands, that are vertical (V), horizontal (H), and diagonal (D), construct CSP constraints, collecting them as $\tau_\theta^{hf} = \{\tau_\theta^V, \tau_\theta^H, \tau_\theta^D\}$.

$$\tau_\theta^V = T_{\theta_V} \cdot p(V_{LR}, V_{ref}), \tag{14}$$

$$\tau_\theta^H = T_{\theta_H} \cdot p(H_{LR}, H_{ref}), \tag{15}$$

$$\tau_\theta^D = T_{\theta_D} \cdot p(D_{LR}, D_{ref}), \tag{16}$$

Finally, sparse high-frequency coefficients are predicted under these constraints:

$$VHD_{SR} \sim \mathcal{N}\left(\mathbb{M}(VHD_{LR}, \tau_\theta^{hf}), sd^2\mathbb{M}\mathbb{M}^\top + \sigma^2 I\right), \tag{17}$$

where $\mathbb{M}$ maps the constrained inputs to predicted details. Algorithm 2 lists the full solver.

Algorithm 2 CSP-WaveDiffUR SDE Solver

Input: Low-resolution image I_{LR}, reference image I_{ref}, pre-trained U-Net s_θ, CSP encoder T_θ, high-frequency restoration module $\mathbb{M}$, current resolution r, target UR resolution R, SR rate k.
Output: Ultra-resolution image I_{UR}.
Compute UR rate $K = R/r$ and number of UR steps $d = K/k$.
Apply Discrete Wavelet Transform (DWT):
$\{A_{LR}, VHD_{LR}\} = DWT(I_{LR})$
Compute CSP constraints:
$\tau_\theta^{lf} = T_\theta \cdot p(I_{ref}, I_{LR}) + \epsilon$
$\tau_\theta^{hf} = T_\theta \cdot p(hf_{ref}, hf_{LR}) + \epsilon$
for $i = 1$ to d (UR steps) **do**
  for $t = T$ to 1 (reverse diffusion process) **do**
    Perform low-frequency wavelet SR:
    $A'_{SR,t-1} = f(A_{SR,t}, s_\theta) + g(A_{SR,t}) \cdot z$
    $A_{SR,t-1} = \alpha \cdot A'_{SR,t-1} + b_i$
  end for
  Perform high-frequency wavelet restoration:
  $VHD_{SR} \sim \mathcal{N}\left(\mathbb{M}(VHD_{LR}, \tau_\theta^{hf}), sd^2\mathbb{M}\mathbb{M}^\top + \sigma^2 I\right)$
  Reconstruct SR image via inverse DWT:
  $I_{SR} = IDWT(\{A_{SR,0}, VHD_{SR}\})$
  if $i \neq d$ **then**
    Update low-resolution image for next iteration:
    $I_{LR} = I_{SR}$
  else
    Assign final UR image:
    $I_{UR} = I_{SR}$
  end if
end for

C. Model architecture

Figure 3 illustrates the overview of the CSP-WaveDiffUR framework. The architecture is deliberately split into two cooperating stages—low-frequency super-resolution (LFSR) and Cross-Scale high-frequency restoration (HFR)—each responsible for a distinct part of the ultra-resolution pipeline. This hierarchical structure first secures a geometrically faithful low-frequency canvas, then injects the missing fine textures.

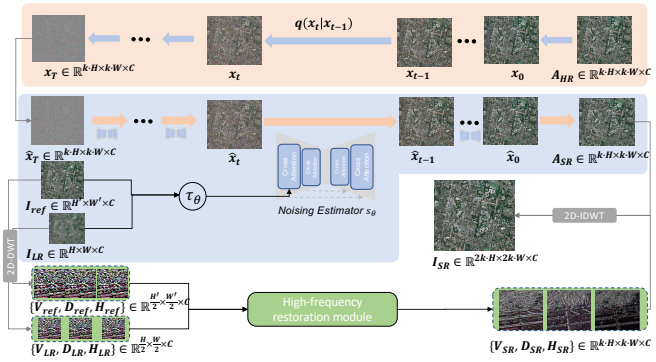

Fig. 3. An overview of the proposed CSP-WaveDiffUR.

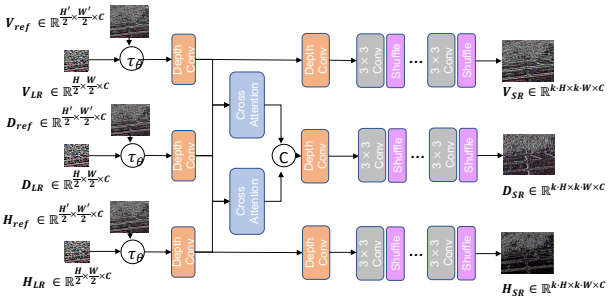

Fig. 5. An illustration of the Cross-Scale High-Frequency Restoration Model (CSHRM).

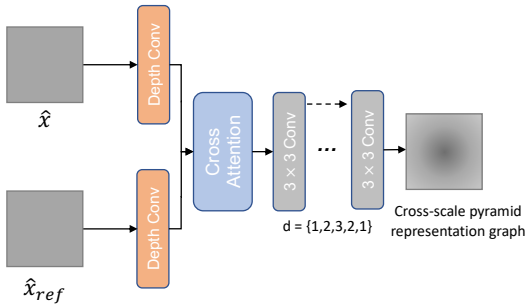

Fig. 4. An illustration of the Cross-Scale Pyramid Encoder.

1) low-frequency super-resolution module: For the LFSR stage, we adopt the Latent Diffusion Model (LDM) U-Net as our baseline generator because of its proven ability to handle conditional probability density transition while keeping memory usage modest. The detailed structure of the LDM refers to Rombach *et al* [48]'s study. To strengthen its cross-scale reasoning, we prepend an external Cross-Scale Pyramid (CSP) encoder. The CSP Encoder is a critical module in the CSP-WaveDiffUR framework, designed to dynamically generate cross-scale constraints for spectral-spatial unmixing. Figure 4 illustrates the structure of the CSP Encoder, which models the spectral-spatial unmixing relationship between a given input ($\hat{x}$) and its corresponding reference ($\hat{x}_{ref}$).

To efficiently extract features from the input, we employ depth-wise separable convolutions [49] in the encoder. The extracted features are then fed into cross-attention layers to compute an intermediate representation tensor. The cross-attention mechanism utilizes information from both $\hat{x}$ and $\hat{x}_{ref}$, effectively modeling the joint probability distribution $p(\hat{x} \mid \hat{x}_{ref})$. The intermediate representation tensor is computed as:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^{\tau}}{\sqrt{d}}\right) \cdot V, \qquad (18)$$

where $Q = W_Q^{(i)} \cdot \phi_i(\hat{x}_{SR,t})$, $K = W_K^{(i)} \cdot \tau_\theta(\hat{x}, \hat{x}_{ref})$, and $V = W_V^{(i)} \cdot \tau_\theta(\hat{x}, \hat{x}_{ref})$. Here, $\phi_i(\hat{x}_{SR,t})$ represents the flattened latent space features generated from a DeepConv block, while $W_Q^{(i)}$, $W_K^{(i)}$, and $W_V^{(i)}$ are learnable projection matrices.

To further enhance the representation of cross-scale spectral-spatial information, we incorporate a progressive dilation residual block (ResBlock) into the encoder. In this block, a sequence of dilated convolutions is applied to expand the receptive field, binding together distant contextual cues while preserving local detail.

2) Cross-Scale High-Frequency Restoration Module: The Cross-Scale High-Frequency Restoration (CSHR) module (Figure 5) in CSP-WaveDiffUR is designed to reconstruct the high-frequency wavelet coefficients $\{V_{SR}, H_{SR}, D_{SR}\}$ from their low-resolution counterparts $\{V_{LR}, H_{LR}, D_{LR}\}$ and reference coefficients $\{V_{ref}, H_{ref}, D_{ref}\}$. This reconstruction enhances the vertical, horizontal, and diagonal details of the target SR image by effectively modeling cross-scale spectral-spatial interactions. The architecture of the CSHR module is illustrated in Figure 5.

The CSHR module consists of the following key steps: (1) Feature Extraction: Depth-wise separable convolutions are applied to efficiently extract features from the input coefficients. (2) Cross-Attention for Detail Enhancement: Two cross-attention layers [50] model the interaction between V and H, augmenting the details in D. (3) Progressive Dilation ResBlock: Inspired by Hai *et al.* [51], a progressive dilation ResBlock is employed. This block consists of a dilation convolution followed by a shuffle layer for the upscaling of high-frequency coefficients.

D. Model Training

In addition to the objective function L_{diff} used for optimizing the diffusion model, we employ a high-frequency realness loss $L_{realness}$, which combines Mean Squared Error (MSE) loss and Total Variation (TV) loss, following a similar approach to Liu *et al.* [52]. This loss function is designed to enhance the reconstruction of high-frequency coefficients.

$$\begin{aligned} L_{realness} = &\lambda_1 \|\{V_{SR}, H_{SR}, D_{SR}\} - \{V_H, H_H, D_H\}\|^2 \\ &+ \lambda_2 TV(\{V_{SR}, H_{SR}, D_{SR}\}), \end{aligned} \qquad (19)$$

where $\lambda_1 = 0.1$ and $\lambda_2 = 2$ are weighting factors for the respective terms. Additionally, we incorporate a consistency loss $L_{consistent}$, which combines $L1$ loss and Structural Similarity Index (SSIM) loss to ensure the fidelity of the reconstructed super-resolution image I_{SR} relative to the ground truth high-resolution image I_H.

$$L_{consistent} = \|I_{SR} - I_H\|_1 + (1 - \text{SSIM}(I_{SR}, I_H)), \quad (20)$$

The total loss for the proposed neural network is given by:

$$L_{total} = L_{diff} + L_{realness} + L_{consistent}, \quad (21)$$

IV. Experiment and Discussion

This section evaluates the performance of the proposed model through several experimental evaluations: (1) comparative analysis against state-of-the-art (SOTA) models on both standard SR and UR tasks, and (2) an ablation study to examine the contribution of individual architectural components.

A. Experimental Setting

1) *Dataset*: To comprehensively evaluate the effectiveness of the proposed methods, we trained and fine-tuned our model on two publicly available datasets: ImageNet 1K [53] and AID [54]. Additionally, we evaluated the model's performance on three public datasets, DOTA drone image [55], DIOR drone image [56], and Houston hyperspectral image [57], as well as a self-designed winter wheat drone-satellite synchronization observation (WWDSSO) multi-spectral dataset [12].

The details of the datasets used for model training, testing, and evaluation are summarized in Table I. In the pre-training phase, we used ImageNet 1K for $64 \times 64 \rightarrow 128 \times 128$ super-resolution tasks, with the development (dev) sets used for validation. We resized the original images to 64×64 (low resolution), 96×96 (reference), and 128×128 (high resolution).

In the fine-tuning phase, we used the AID dataset for $64 \times 64 \rightarrow 256 \times 256$ remote sensing super-resolution tasks. The AID training set consisted of 8000 randomly selected images, while the remaining 2000 images were used as the testing set. Each image was processed into three corresponding resolutions: 64×64, 96×96, and 256×256.

For model evaluation, we selected subsets of images from the publicly available DOTA, DIOR, and Houston datasets and self-collected WWDSSO datasets, as shown in Table I. During the data pre-processing phase, we applied bicubic interpolation for image degradation. For real-world analysis, we used the self-collected WWDSSO dataset. This dataset comprised 300 drone-satellite synchronous observation pairs, including Landsat-8 images with a 30m resolution (as low-resolution input), Sentinel-2 images with a 10m resolution (as mid-resolution reference), and multispectral drone images with a 0.23m resolution (as ground truth). We evaluated super-resolution performance at three scale factors: $\times 10$, $\times 20$, and $\times 100$. Simulated degradation was applied to the ground-truth drone images for scale factors ranging from $\times 2$ to $\times 128$ super-resolution.

TABLE I
THE DETAILS OF THE DATASETS FOR MODEL TRAINING, TESTING, AND EVALUATION.

Data type	ImageNet	AID	DOTA	DIOR	Houston	WWDSSO
Scale for	$\times 2$	$\times 2$, $\times 4$	$\times 2$, $\times 4$	$\times 2$, $\times 4$	$\times 4$ to $\times 64$	$\times 4$ to $\times 128$
Pre-train	1,281,167					
Fine-tune		8,000				
Testing	500,000	2,000	300	300		
Evaluation			180	300	129	300

2) *Evaluation Metrics*: To comprehensively assess the performance of the super-resolution model, we employed seven evaluation metrics. These included three full-reference metrics that measured similarity between the super-resolution and ground-truth images: Fréchet Inception Distance (FID) [58], the widely used Peak Signal-to-Noise Ratio (PSNR) [12], [59], and Structural Similarity Index (SSIM) [60]. Among these, FID was extensively used in evaluating the generative quality of the model, as it improved upon the Inception Score (IS) [7] by directly measuring feature-level distances without relying on a classifier.

To assess pixel-wise spectral fidelity, we employed two additional metrics: Spectral Angle Mapper (SAM) [61] and Spectral Reconstruction Error (SRE) [12]. These metrics computed the average spectral angle and reconstruction error between super-resolution images and ground-truth data, ensuring spectral consistency.

Furthermore, we incorporated two reference-free metrics: the Natural Image Quality Evaluator (NIQE) [62], which quantified perceptual quality, and Average Gradient (AG) [7], which evaluated the preservation of high-frequency details.

These metrics provided a holistic evaluation of the model's performance by addressing: Quantitative accuracy (PSNR, SSIM); Perceptual quality (FID, NIQE); Spectral consistency (SAM, SRE); Sharpness and detail preservation(AG).

3) *Comparative Methods*: To assess the performance of the proposed model, we conducted a comparative analysis. This analysis included our model and several state-of-the-art (SOTA) super-resolution approaches, including LE-GAN [12], ViT-ISRGAN [63], DiffuseVAE [17], EDIP-Net [57], SR3 [32], IRSDE [33], EDiffSR [7], and LWTDM [8]. These SOTA SR models are dominant techniques in the field and represent diverse methodologies. Specifically, LE-GAN and ViT-ISRGAN represent GAN-based methods, while DiffuseVAE and EDIP-Net correspond to VAE-based image super-resolution methods. Conversely, SR3 and IRSDE are cutting-edge diffusion-based models for natural image super-resolution. EDiffSR and LWTDM are diffusion-based models specifically designed for remote sensing image super-resolution. All comparative models were fine-tuned on the AID training dataset following the configurations specified in their official implementations, ensuring a fair and consistent comparison.

4) *Implementation Details*: We used a high-end GPU workstation with one NVIDIA A100 Tensor Core GPU and 40 GB of memory to run the algorithms in PyTorch. Training used the Adam optimizer with a learning rate of 0.001 and a batch size of 32.

TABLE II
A QUANTITATIVE COMPARISON OF SOTA SR MODELS (LE-GAN [12], VIT-ISRGAN [63], DIFFUSEVAE [17], EDIP-NET [57], SR3 [32], IRSDE [33], EDIFFSR [7], AND LWTDM [8]) PRE-TRAINED ON THE IMAGENET DATASET (×2) AND FINE-TUNED ON THE AID DATASET (×2, ×4) IN TERMS OF FID, PSNR, SSIM, SAM, SRE, NIQE, AND AG. THE TESTS WERE CONDUCTED ON THE DEV SPLIT TEST DATA. THE BEST VALUES ARE HIGHLIGHTED.

ImageNet									
Category	Method	Scale	FID ↓	PSNR↑	SSIM↑	SAM ↓	SRE↓	NIQE ↓	AG ↑
GAN	LE-GAN	×2	48.39	48.27	0.85	**5.61**	**5.16**	12.82	4.51
	ViT-ISRGAN		53.26	48.02	0.84	6.97	8.35	12.24	4.81
VAE	DiffuseVAE		47.15	45.61	0.8	7.66	9.04	13.27	4.98
	EDIP-Net		47.44	45.49	0.84	7.84	8.28	12.84	5.12
Diffusion	SR3		40.52	50.47	**0.95**	6.55	7.32	10.48	4.82
	IRSDE		45.03	48.56	0.85	6.62	7.94	12.42	5.36
	EDiffSR		**40.41**	49.85	0.88	6.68	7.17	11.04	5.62
	LWTDM		43.98	46.3	0.77	7.8	9.07	14.27	4.33
	Proposed		40.45	**52.97**	**0.95**	5.79	5.65	**9.86**	**5.85**
AID									
Category	Method	Scale	FID ↓	PSNR↑	SSIM↑	SAM ↓	SRE↓	NIQE ↓	AG ↑
GAN	LE-GAN	×2	53.13	39.8	0.71	8.09	**7.68**	14.3	3.9
	ViT-ISRGAN		55.2	39.62	0.7	9.66	10.71	16.08	3.91
VAE	DiffuseVAE		59.76	37.7	0.67	10.39	11.46	17.01	3.73
	EDIP-Net		60.21	37.22	0.7	10.58	11.37	17.06	3.65
Diffusion	SR3		51.43	40.8	0.67	10.07	10.31	13.2	3.98
	IRSDE		53.8	40.41	0.71	9.35	10.46	15.7	4.06
	EDiffSR		51.49	**41.49**	0.73	9.11	9.99	14.98	4.07
	LWTDM		55.06	39.56	0.64	10.72	12.03	18.04	3.65
	Proposed		**50.9**	41.48	**0.75**	**8.07**	7.71	**12.03**	**4.24**
GAN	LE-GAN	×4	56.19	35.14	0.7	10.14	9.12	16.32	3.46
	ViT-ISRGAN		61.05	35.64	0.63	11.31	12.75	18.67	2.82
VAE	DiffuseVAE		66.04	33.82	0.6	12.01	13.24	19.26	3.35
	EDIP-Net		66.25	32.97	0.63	11.79	13.12	19.04	3.17
Diffusion	SR3		66.02	35.85	0.6	11.51	11.84	14.86	3.4
	IRSDE		59.58	35.75	0.64	11.22	11.82	17.36	2.98
	EDiffSR		56.67	37.17	0.66	10.06	11.05	16.8	3.55
	LWTDM		67.34	32.23	0.58	12.63	14.11	20.11	2.52
	Proposed		**53.18**	**38.64**	**0.71**	**8.97**	**8.47**	**13.75**	**4.14**

TABLE III
A QUANTITATIVE COMPARISON OF THE GENERALIZATION CAPABILITY OF SOTA SR MODELS (LE-GAN [12], VIT-ISRGAN [63], DIFFUSEVAE [17], EDIP-NET [57], SR3 [32], IRSDE [33], EDIFFSR [7], AND LWTDM [8]) ON INDEPENDENT DOTA, DIOR, HOUSTON, AND WWDSSO DATASETS FOR ×4 SUPER-RESOLUTION TASKS IN TERMS OF FID, PSNR, SSIM, SAM, SRE, NIQE, AND AG. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

DOTA									
Category	Method	Scale	FID ↓	PSNR↑	SSIM↑	SAM ↓	SRE↓	NIQE ↓	AG ↑
GAN	LE-GAN	×4	26.34	21.31	0.78	**6.48**	7.04	14.68	2.14
	ViT-ISRGAN		26.89	21.69	0.73	9.81	10.78	17.43	3.48
VAE	DiffuseVAE		29.65	20.27	0.81	10.99	13.27	18.67	2.28
	EDIP-Net		29.46	20.26	0.8	12.51	12.73	18.22	2.09
Diffusion	SR3		24.93	21.82	0.79	11.51	10.81	14.42	2.73
	IRSDE		26.82	21.9	0.7	11.17	11.96	16.63	3.07
	EDiffSR		24.92	22.66	0.71	9.21	11.03	16.77	2.58
	LWTDM		27.48	21.16	0.74	11.65	13.22	19.97	1.9
	Proposed		**20.63**	**28.53**	**0.93**	6.58	**6.61**	**10.18**	**6.23**
DOTA									
GAN	LE-GAN	×4	26.72	19.55	0.72	11.86	11.41	19.37	2.06
	ViT-ISRGAN		31.29	20.54	0.67	14.06	14.4	20.78	3.45
VAE	DiffuseVAE		34.52	20.15	0.72	14.57	16.3	23.04	2.24
	EDIP-Net		32.98	17.94	0.71	15.27	15.62	21.65	2.04
Diffusion	SR3		27.44	20.43	0.72	13.07	15.46	18.05	2.69
	IRSDE		27.48	19.62	0.62	13.15	16.21	20.96	3.06
	EDiffSR		27.75	22.34	0.68	14.22	13.22	20.84	2.55
	LWTDM		30.31	19.87	0.59	15.2	14.15	22.34	1.87
	Proposed		**22.65**	**28.21**	**0.91**	**7.27**	**7.82**	**11.98**	**6.17**
Houston									
Category	Method	Scale	FID ↓	PSNR↑	SSIM↑	SAM ↓	SRE↓	NIQE ↓	AG ↑
GAN	LE-GAN	×4	26.09	22.71	0.79	10.11	11.75	18.97	5.41
	ViT-ISRGAN		30.56	18.46	0.65	11.17	18.06	18.66	6.37
VAE	DiffuseVAE		30.59	17.04	0.62	12.67	24.37	20.51	2.62
	EDIP-Net		31.49	22.7	0.64	11.26	20.88	19.57	2.66
Diffusion	SR3		27.7	23.75	0.64	13.42	18.7	18.52	3.2
	IRSDE		30.81	22.84	0.68	14.99	18.02	16.36	4.05
	EDiffSR		28.32	24.91	0.63	13.73	18.44	15.92	5.07
	LWTDM		29.35	19.05	0.65	12.78	18.09	19.78	3.25
	Proposed		**21.04**	**27.76**	**0.87**	**9.71**	**8.37**	**15.12**	**7.89**
WWDSSO									
GAN	LE-GAN	×4	33.62	19.4	0.62	12.69	12.69	19.77	2.04
	ViT-ISRGAN		32.97	18.65	0.59	18.38	16.39	23.13	3.44
VAE	DiffuseVAE		36.89	17.78	0.54	13.24	19.51	19.66	2.15
	EDIP-Net		35.12	17.26	0.56	16.19	15.55	22.37	1.97
Diffusion	SR3		32.23	18.73	0.58	15.61	14.62	19.11	2.59
	IRSDE		32.91	17.15	0.58	15.09	13.63	21.15	2.97
	EDiffSR		27.41	18.35	0.57	14.86	15.92	22.1	2.56
	LWTDM		31.8	18.63	0.53	14.32	15.17	24.39	1.79
	Proposed		**24.63**	**27.71**	**0.82**	**10.04**	**9.71**	**14.66**	**5.92**

B. Model Evaluation on SR Tasks

The average FID, PSNR, SSIM, SAM, NIQE, and AG values for the ImageNet and AID test sets are shown in Table II. On the ImageNet dataset, the proposed model achieves top-tier performance, ranking first or exhibiting negligible differences ($< 0.1\%$) from leading SOTA models (e.g., SR3 and EDiffSR) in FID, PSNR, and SSIM. Notably, the proposed model surpasses diffusion-based models, achieving a 13.12% higher SAM score than SR3. It also obtains the best reference-free metrics (NIQE and AG), underscoring its structural integrity in low-frequency components. Similar trends are observed on the AID dataset. The proposed model matches EDiffSR in FID, PSNR, and SSIM, is comparable to LE-GAN in SAM and SRE, and outperforms in NIQE and AG.

Additional results for assessing the generalization capacity of the models on $\times 4$ SR on independent RGB (DOTA and DIOR), hyperspectral (Houston), and multispectral (WWDSSO) datasets are shown in Table III, where the highest performances are highlighted in bold. The evaluation results indicate that the proposed model ranks first in 6 out of 7 metrics on the DOTA dataset and achieves the highest score in all metrics for the DIOR, Houston, and WWDSSO datasets. These findings confirm the proposed model's excellent generalization capacity.

A visual comparison between the proposed model and the SOTA SR models is also conducted. Figure 6 demonstrates a comparison of $\times 2$ super-resolution results ($64 \times 64 \rightarrow 128 \times 128$) on the AID, DOTA, and DIOR test sets. Our proposed model consistently generates more realistic and detailed reconstruction closely resembling the ground truth image. In the case of the "$farmland_1 45$" image from the AID test set, the proposed model captures fine textures in farmland patches. The VAE-based models like DiffuseVAE and LWTDM show over-smoothed regions, losing critical texture details. In the case of the "$P2541$" image from the DOTA test set, the grass area reconstructed by the proposed model aligns more accurately with the ground truth, but competing models (e.g., DiffuseVAE) exhibit blurred details, reflecting weaker spatial generalization. In the case of the "21778" image from the DIOR test set, the proposed model restores clear spatial details of densely built-up areas (as highlighted in the zoomed-in frames), outperforming other models like DiffuseVAE, IRSDE, and LWTDM, which produce noticeable blurs.

C. Evaluation for UR Tasks

In this section, we present the evaluation results of the proposed WaveDiffUR SDE solver for UR tasks. In the DOTA, DIOR, and Houston testing sets, the highest-resolution data was treated as ground truth, and lower-resolution data was generated using bicubic downsampling. For UR resolutions beyond the highest available ground truth, pseudo-ground-truth images were generated by upsampling the ground truth to match the UR results for evaluation. PSNR and SRE were employed to evaluate spatial fidelity and spectral consistency, respectively. The results are presented in Table IV.

The proposed model achieved the highest PSNR and SRE scores across all independent datasets and for all evaluated

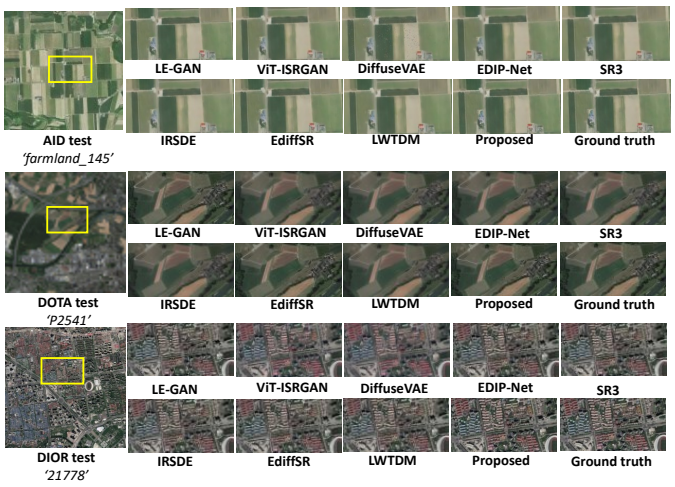

Fig. 6. $\times 2$ visual comparisons of the proposed model and the SOTA SR models (LE-GAN [12], ViT-ISRGAN [63], DiffuseVAE [17], EDIP-Net [57], SR3 [32], IRSDE [33], EDiffSR [7], and LWTDM [8]) on AID, DOTA, and DIOR test sets. The square image at the left side is the input 64×64 image, and the rectangular frames indicate the zoomed-in view of the $\times 2$ super-resolution results for a better view.

upscaling factors. It significantly outperformed its competitors, with an approximate 43% improvement in PSNR and SRE for large-scale ($> \times 32$) UR. At extreme magnifications (e.g., $\times 128$), it achieved up to $2\times$ improvement in PSNR and SRE. This is attributed to the Cross-Scale Pyramid constraint in the proposed model, which effectively guides diffusion-denoising inference, enabling accurate reconstruction of both spectral and spatial details in UR tasks.

Figure 7 provides a comparative analysis of visual results when increasing the UR scale factor from $\times 16$ to $\times 128$. The analysis highlights the challenges and performance degradation of existing methods when increasing the scale factor. Most SOTA models exhibit noticeable degradation in UR performance. For example, in the "11834" image from the DIOR test, EDIP-Net and DiffuseAVE produce over-smoothed outputs. In the "$P0615$" image from the DOTA test, the bridges reconstructed by DiffuseAVE, IRSDE, and LWTDM are plagued with textural blurring. In the Houston test, LWTDM introduces noticeable color artefacts. In contrast, our proposed model leverages the cross-scale pyramid architecture of the proposed model in predicting contextual prior to recover fine-grained texture, enhancing the performance of diffusion models in UR tasks. Such as the details of the bridge in "$P0615$" image from the DOTA test, and the details of the building in the Houston test.

D. Ablation Studies

In this section, we present extensive experiments to demonstrate the effectiveness of each component within our self-cascade model.

1) *Component analysis*: To evaluate the effectiveness of each component in the proposed self-cascade model, we conducted an ablation study by removing three key elements from the proposed model: CSP conditions, high-frequency restoration, and the self-cascade structure, one by one. This produced four simplified models: Baseline (i.e. Latent Diffusion U-Net [48]), Model-1 (adding CSP constraints only), Model-2 (adding wavelet domain only), Model-3 (adding CSP constraints and wavelet domain), and the proposed model. Table V summarizes the model configurations and performance evaluation results in terms of FID, SRE, and NIQE. As shown in Table V, the non-wavelet diffusion U-Net (baseline) failed to converge in terms of FID, SRE, and NIQE, indicating the baseline generated results suffer from severe blur in the spectral and spatial domain. In comparison, Model-1 generated better fidelity in FID (63.62 vs. 81.28), indicating the wavelet decomposition helps constrain the solution in UR process. Model-2 produced better results in spectral details (SRE: 19.15 vs. 27.74), indicating that high-frequency restoration improves spectral consistency.

When both CSP constraints and the high-frequency restoration module are combined into the Model-3 model, the FID and SRE metrics improve substantially. The proposed embedding with self-cascade strategy requires no additional external parameters, as the same model is reused without fine-tuning. This integration yields significant improvements across all

TABLE IV
A QUANTITATIVE COMPARISON OF MODEL PERFORMANCE IN TERMS OF PSNR AND SRE METRICS OF BACKBONE MODELS (SCALING FROM $\times 8$ TO $\times 128$) USING THE INDEPENDENT DATASETS DOTA, DIOR, HOUSTON, AND WWDSSO. OUTPUTS MARKED WITH AN ASTERISK (*) INDICATE EVALUATIONS PERFORMED WITH PSEUDO-GROUND-TRUTH IMAGES, WHERE THE UR IMAGE RESOLUTION EXCEEDS THE GROUND TRUTH RESOLUTION.

Method	DOTA Scale 8	16	32	64*	128*
LE-GAN	25.56/8.28	23.91/12.12	21.56/13.87	15.37/16.93	16.51/17.63
ViT-ISRGAN	24.81/9.28	24.71/9.72	19.93/12.02	14.21/16.69	14.31/21.13
DiffuseVAE	25.78/7.82	24.41/10.66	21.34/12.2	16.38/14.43	13.61/19.67
EDIP-Net	26.61/9.98	25.14/10.63	20.34/14.03	16.61/20.99	13.94/26.44
SR3	24.92/10.25	23.38/10.74	20.98/12.74	15.32/22.26	13.58/26.75
IRSDE	25.32/15.31	21.95/14.32	16.35/20.32	14.92/24.12	11.95/34.33
EDiffSR	28.18/7.88	26.21/9.83	21.21/12.83	18.18/17.89	16.21/19.84
LWTDM	23.99/15.62	19.42/16.92	16.62/23.32	15.79/24.43	9.22/37.93
Proposed	28.31/7.37	27.32/8.23	23.32/12.23	20.31/14.38	17.32/15.24

Method	DIOR Scale 8	16	32*	64*	128*
LE-GAN	21.56/11.88	19.31/11.92	18.56/14.87	11.77/18.53	12.31/19.43
ViT-ISRGAN	23.19/9.78	22.71/11.42	18.35/14.74	13.81/15.99	9.92/21.53
DiffuseVAE	23.33/10.68	21.42/11.86	19.81/12.89	15.19/16.52	11.05/20.15
EDIP-Net	23.82/11.78	20.62/12.54	15.91/14.98	15.51/21.33	14.19/24.63
SR3	22.59/10.82	19.59/13.71	17.29/15.12	14.01/18.54	12.15/23.71
IRSDE	21.38/9.83	18.48/13.86	15.42/17.82	11.01/22.41	6.32/27.53
EDiffSR	23.21/9.81	21.92/10.32	18.84/14.32	14.22/17.12	10.36/21.96
LWTDM	21.03/11.11	17.72/12.23	15.72/16.91	12.91/19.01	9.31/22.88
Proposed	24.92/8.87	23.31/8.72	22.33/12.21	17.79/15.42	13.32/18.85

Method	Houston Scale 8	16	32	64	128*
LE-GAN	24.96/9.28	21.51/10.52	17.66/13.67	18.97/16.73	17.01/16.83
ViT-ISRGAN	32.61/11.67	30.79/13.82	27.39/18.34	23.81/22.19	18.41/25.31
DiffuseVAE	29.71/9.36	33.19/11.78	28.99/14.12	20.85/21.35	21.42/22.57
EDIP-Net	34.32/14.14	33.31/14.61	28.01/15.13	26.19/20.43	21.72/30.8
SR3	33.29/11.31	27.19/14.72	22.49/17.01	20.65/19.31	14.65/28.71
IRSDE	26.78/17.86	25.92/21.52	16.71/28.39	14.18/33.73	10.88/32.82
EDiffSR	35.42/10.19	32.34/10.32	24.72/14.52	21.36/17.62	17.63/20.81
LWTDM	28.22/18.03	26.27/22.91	19.61/26.29	12.58/29.28	11.91/34.45
Proposed	37.69/7.35	32.83/8.46	29.79/12.75	26.32/14.65	23.98/15.45

Method	WWDSSO Scale 8	16	32	64	128
LE-GAN	28.36/7.48	26.91/9.92	20.36/11.67	17.37/15.73	18.51/17.03
ViT-ISRGAN	27.71/16.47	24.39/19.42	22.59/21.34	19.01/25.99	14.21/30.31
DiffuseVAE	28.61/12.76	26.49/15.38	23.39/18.72	18.95/22.75	15.52/26.17
EDIP-Net	30.12/16.94	25.71/18.41	23.61/18.53	17.09/25.63	13.82/33.2
SR3	29.39/13.51	26.99/17.12	21.79/17.21	17.35/21.51	14.95/28.91
IRSDE	22.88/23.26	20.62/25.32	13.41/32.19	8.48/36.53	6.68/36.62
EDiffSR	31.92/10.39	28.84/12.32	24.22/15.12	20.36/19.62	16.13/21.21
LWTDM	28.32/19.03	25.97/25.11	22.51/26.69	15.38/31.88	14.11/32.85
Proposed	34.19/8.75	31.33/9.26	27.79/13.95	25.32/14.85	22.48/16.85

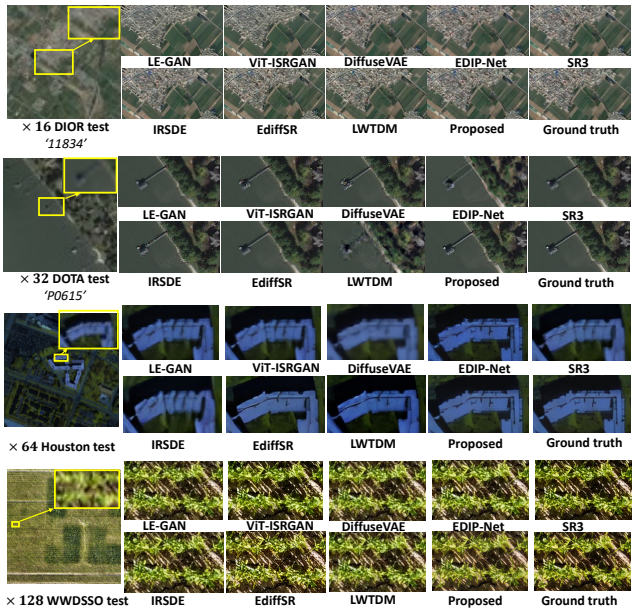

Fig. 7. The visual comparisons with SOTA SR models (LE-GAN [12], ViT-ISRGAN [63], DiffuseVAE [17], EDIP-Net [57], SR3 [32], IRSDE [33], EDiffSR [7], and LWTDM [8]) on $\times 16$ DIOR, $\times 32$ DOTA, $\times 64$ Houston (R:46, G:30, B:14) and $\times 128$ WWDSSO (R:4, G:3, B:2) test set. The square image at the left side is the input 64×64 image, and the rectangular frames indicate the zoomed-in view of the UR results for a better view.

TABLE V
ABLATION ANALYSIS OF THE PROPOSED METHODOLOGY IN THE $\times 8$ UR TASK AS AN EXAMPLE.

Model	CSP constraints	Wavelet domain	Self-cascade	Param.(M)	FID	SRE	NIQE
Baseline [48]				18.84	81.28	27.74	18.89
Model-1	✓			24.82	63.62	26.49	16.82
Model-2		✓		22.14	87.82	19.15	15.52
Model-3	✓	✓		29.12	59.15	16.24	14.33
Proposed	✓	✓	✓	29.12	54.82	9.81	14.03

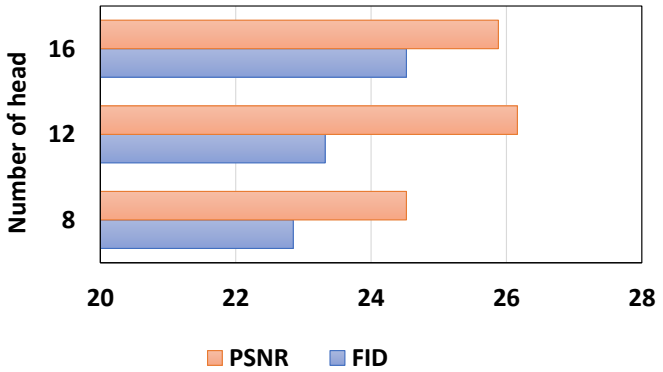

Fig. 8. Ablation analysis of CSP constraint conditions with different numbers of heads in cross-attention blocks in terms of FID and SRE.

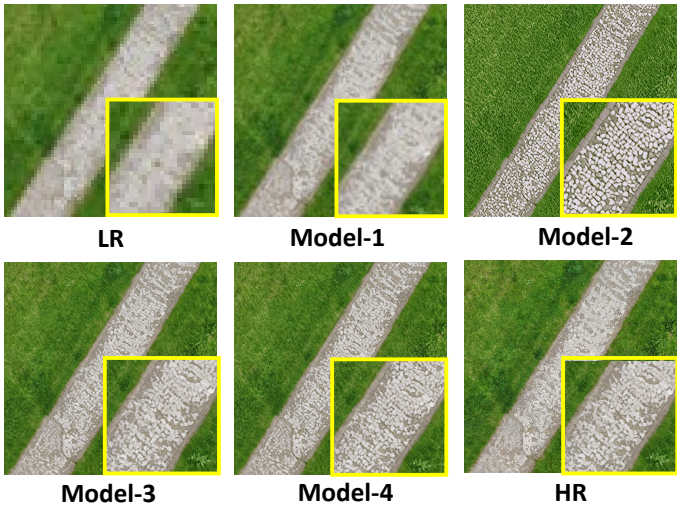

Fig. 9. Visual ablation analysis of the $\times 128$ UR results from the model configurations listed in Table V.

metrics, including FID, SRE, and NIQE.

These findings demonstrate the effectiveness of the proposed methodology in enhancing large-scale image upscaling. Moreover, the low-complexity design of the components ensures that the self-cascade architecture remains both efficient and powerful.

2) *Effectiveness of CSP constraints*: We analyzed the impact of varying the number of heads in the cross-attention blocks for modeling CSP constraint conditions. As shown in Figure 8, the proposed model achieves slightly better FID performance with 16 heads compared to 12 heads, while the highest PSNR results are obtained with 12 heads. To balance model size and performance effectively, we set the default number of heads to 12.

3) *Effectiveness of high-frequency restoration*: To demonstrate the capability of high-frequency restoration in recovering fine details for accurate UR reconstruction, we present a visual comparison in Figure 9.

Comparing the UR outputs of Model-1 (cross-scale condition only) and Model-2 (high-frequency predictor only), it is evident that Model-2 produces overly sharpened details compared to Model-1. Comparing the UR results of Model-3 and proposed model, the combination of the high-frequency predictor and cross-scale condition achieves a superior balance between high-frequency and low-frequency components. This is particularly noticeable in features such as road edges and surface textures. These observations emphasize that the high-frequency predictor significantly enhances UR performance by predicting enriched high-frequency details with additional priors, effectively improving overall reconstruction quality.

4) *Effectiveness of self-cascade strategy*: Figure 10 illustrates a quantitative comparison of model performance between traditional one-step fine-tuning and self-cascade strategy with different backbone models in terms of NIQE and AG metrics. Performance degradation is observed across all models in large-scale UR tasks, but the self-cascade strategy with different backbone models demonstrates robust performance. The proposed approach achieves the best performance in NIQE and AG, with only 11.8% and 19.1% performance degradation in NIQE and AG, respectively, as the magnification scale increases from $\times 4$ to $\times 128$. These observations indicate the proposed WaveDiffUR architecture with self-cascade strategy is compatible with existing SR models to handle the ill-pose problem in the UR task.

To provide an intuitive comparison, we present a visual

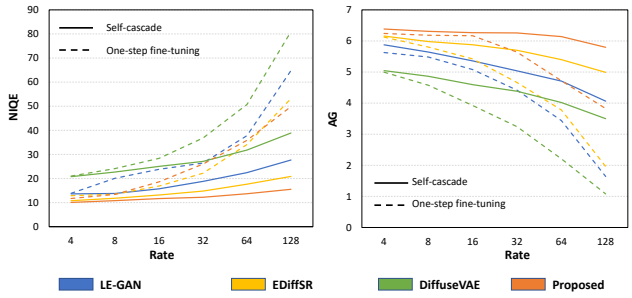

Fig. 10. A comparison of model performance between WaveDiffUR-based models with different backbone models (i.e., LE-GAN, EDiffSR, and DiffuseVAE) in the traditional one-step fine-tuning and self-cascade strategy for UR tasks in terms of NIQE and AG.

analysis of the traditional one-step fine-tuning and self-cascade strategy in Figure 11. We observe that the self-cascade-based UR images recover more accurate details (e.g., white lines on the bay and embankment in simulated evaluations, or roads in real-world evaluations). In contrast, the one-step-based UR images tend to introduce exaggerated sharpening with pseudo-details. A possible explanation is the accumulation of cross-scale biases during the one-step transition, which hampers the realistic reconstruction of high-frequency components, leading to overly sharp and unnatural details.

5) *Model efficiency*: To evaluate the efficiency of the self-cascade strategy, we compared the parameters, VRAM usage, training time, number of FLOPs, and inference times of the models on one NVIDIA A100 Tensor Core GPU and 40 GB of memory. As shown in Figure 12, while the proposed baseline model is not the lightest among existing DPM-based SR models in terms of parameters, VRAM usage and the number of FLOPs, it remains highly competitive in terms of efficiency.

Notably, the proposed model within the self-cascade UR architecture demonstrates faster inference speeds compared to its competitors, especially for large-scale UR tasks ($> \times 16$). This efficiency gain is primarily due to the wavelet transformation integrated into our model. By compressing the input data and processing high-frequency and low-frequency components in parallel at each UR scale, the wavelet transformation reduces memory requirements for storing intermediate feature maps and decreases the number of convolutional operations, thereby lowering computational costs. These optimizations enable the self-cascade architecture to outperform other approaches in computational efficiency, making our method more practical and scalable for real-world applications.

V. Conclusions and Future Work

This study introduces the WaveDiffUR architecture, addressing the challenges of remote sensing SR and UR tasks. The proposed CSP-WaveDiffUR model achieves superior performance compared to SOTA methods by incorporating CSP constraint conditions based on cross-scale spectral-spatial unmixing rules. This approach effectively mitigates degradation in accuracy, perceptual quality, spectral consistency, and detail sharpness. CSP-WaveDiffUR achieves up to threefold

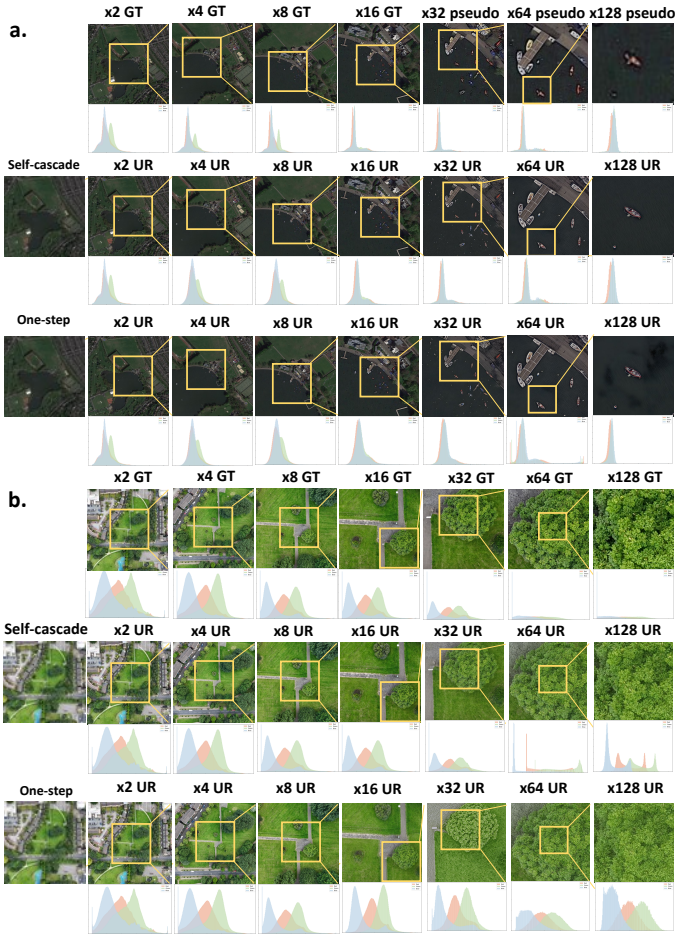

Fig. 11. Visual comparison of UR images generated by the one-step and self-cascade approaches, highlighting spatial fidelity and channel consistency on (a) the DIOR dataset and (b) the WWDSSO dataset. Zoom-in views within the yellow boxes offer enhanced visual detail.

improvement in PSNR and twofold reduction in SRE at extreme magnifications (e.g., $\times 128$), while maintaining leading performance in NIQE and AG metrics.

Despite its advantages, the method relies on high-quality LR and reference image pairs and struggles with degradation variability across different systems. Future work will focus on reducing dependency on reference images for blind SR tasks and enhancing adaptability to diverse degradation patterns. Additionally, we aim to improve the model's generalization ability by incorporating self-supervised learning techniques and domain adaptation strategies, enabling robust performance across diverse imaging conditions. Moreover, integrating real-time processing capabilities will be a key focus, facilitating deployment in time-sensitive applications such as disaster response and environmental monitoring. These improvements aim to extend the framework's applicability and robustness, with potential transformative impacts on environmental monitoring, urban planning, disaster response, and precision agriculture. We will release the source code on GitHub upon publication with this link: https://github.com/nedvede/WaveDiffUR. We hope this will encourage further research in remote sensing ultra-resolution.

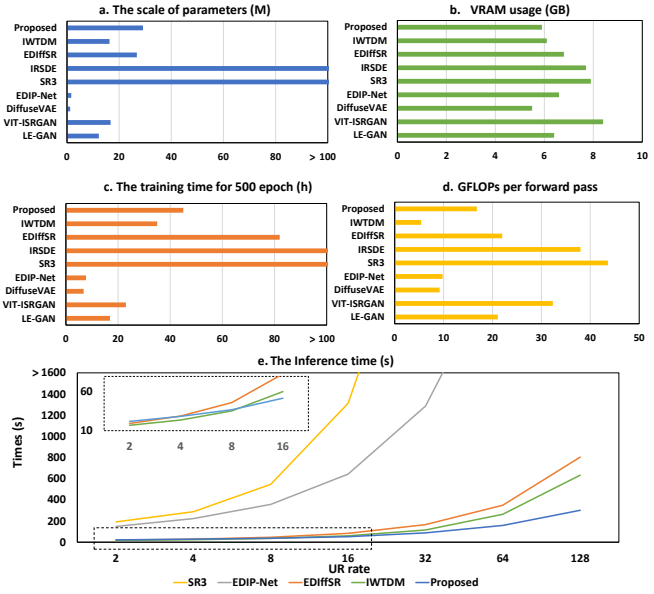

Fig. 12. Efficiency comparison of SR models: (a) number of parameters (M), (b) VRAM usage (GB), (c) training time for 500 epochs on a single NVIDIA A100 GPU (hours), (d) GFLOPs per forward pass, and (e) inference time (s) across varying upscaling rates (UR). The proposed model demonstrates competitive efficiency and significantly faster inference, especially at higher UR scales (e.g., ×16 and above.

REFERENCES

[1] G. Vivone, L.-J. Deng, S. Deng, D. Hong, M. Jiang, C. Li, W. Li, H. Shen, X. Wu, J.-L. Xiao *et al.*, "Deep learning in remote sensing image fusion: Methods, protocols, data, and future perspectives," *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, pp. 269 – 310, 2024.

[2] X. Wang, J. Yi, J. Guo, Y. Song, J. Lyu, J. Xu, W. Yan, J. Zhao, Q. Cai, and H. Min, "A review of image super-resolution approaches based on deep learning and applications in remote sensing," *Remote Sensing*, vol. 14, no. 21, p. 5423, 2022.

[3] Z.-C. Wu, Y.-J. Li, T.-Z. Huang, L.-J. Deng, and G. Vivone, "Crodosr: Tensor cross-domain rank for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing,*, vol. 62, 2024.

[4] S. Chen, L. Zhang, and L. Zhang, "Cross-scope spatial-spectral information aggregation for hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 33, pp. 5878–5891, 2024.

[5] Q. Liu, X. Meng, F. Shao, and S. Li, "Supervised-unsupervised combined deep convolutional neural networks for high-fidelity pansharpening," *Information Fusion*, vol. 89, pp. 292–304, 2023.

[6] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[7] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2023.

[8] T. An, B. Xue, C. Huo, S. Xiang, and C. Pan, "Efficient remote sensing image super-resolution via lightweight diffusion models," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, 2024.

[9] D. He, Q. Shi, J. Xue, P. M. Atkinson, and X. Liu, "Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation," *Remote Sensing of Environment*, vol. 299, p. 113884, 2023.

[10] Y. Shi, L. Han, A. Kleerekoper, S. Chang, and T. Hu, "Novel cropdocnet model for automated potato late blight disease detection from unmanned aerial vehicle-based hyperspectral imagery," *Remote Sensing*, vol. 14, no. 02, p. 396, 2022.

[11] H. Fu, F. Peng, X. Li, Y. Li, X. Wang, and H. Ma, "Continuous optical zooming: A benchmark for arbitrary-scale image super-resolution in real world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3035–3044.

[12] Y. Shi, L. Han, L. Han, S. Chang, T. Hu, and D. Dancey, "A latent encoder coupled generative adversarial network (le-gan) for efficient hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[13] X. Zhu, L. Zhang, L. Zhang, X. Liu, Y. Shen, and S. Zhao, "Gan-based image super-resolution with a novel quality loss," *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 5217429, 2020.

[14] J. Song, H. Yi, W. Xu, X. Li, B. Li, and Y. Liu, "Esrgan-dp: Enhanced super-resolution generative adversarial network with adaptive dual perceptual loss," *Heliyon*, vol. 9, no. 4, 2023.

[15] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Photo-realistic image super-resolution via variational autoencoders," *IEEE Transactions on Circuits and Systems for video Technology*, vol. 31, no. 4, pp. 1351–1365, 2020.

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

[17] 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.

[18] Y. Liu, J. Yue, S. Xia, P. Ghamisi, W. Xie, and L. Fang, "Diffusion models meet remote sensing: Principles, methods, and perspectives," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–22, 2024.

[19] Z. Yu, M. Y. I. Idris, and P. Wang, "A diffusion-based framework for terrain-aware remote sensing image reconstruction," 2025. [Online]. Available: https://arxiv.org/abs/2504.12112

[20] Y. Zhang and M. He, "Multi-spectral and hyperspectral image fusion using 3-d wavelet transform," *Journal of electronics (China)*, vol. 24, pp. 218–224, 2007.

[21] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, and J. Chanussot, "Hyperspectral pansharpening using deep prior and dual attention residual network," *IEEE transactions on geoscience and remote sensing*, vol. 58, no. 11, pp. 8059–8076, 2020.

[22] J. Li, K. Zheng, W. Liu, Z. Li, H. Yu, and L. Ni, "Model-guided coarse-to-fine fusion network for unsupervised hyperspectral image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[23] J. Li, K. Zheng, Z. Li, L. Gao, and X. Jia, "X-shaped interactive autoencoders with cross-modality mutual learning for unsupervised hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[24] X. Zheng, R. Feng, J. Fan, W. Han, S. Yu, and J. Chen, "Msisr-stf: Spatiotemporal fusion via multilevel single-image super-resolution," *Remote Sensing*, vol. 15, no. 24, p. 5675, 2023.

[25] L. Gao, J. Li, K. Zheng, and X. Jia, "Enhanced autoencoders with attention-embedded degradation learning for unsupervised hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[26] L. Chen, H. Liu, M. Yang, Y. Qian, Z. Xiao, and X. Zhong, "Remote sensing image super-resolution via residual aggregation and split attentional fusion network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 9546–9556, 2021.

[27] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, "Real-world single image super-resolution: A brief review," *Information Fusion*, vol. 79, pp. 124–145, 2022.

[28] P. Behjati, P. Rodriguez, C. Fernández, I. Hupont, A. Mehri, and J. Gonzàlez, "Single image super-resolution based on directional variance attention network," *Pattern Recognition*, vol. 133, p. 108997, 2023.

[29] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3d full convolutional neural network," *Remote Sensing*, vol. 9, no. 11, p. 1139, 2017.

[30] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1082–1096, 2020.

[31] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, and W. Xu, "Improved srgan for remote sensing image super-resolution across locations and sensors," *Remote Sensing*, vol. 12, no. 8, p. 1263, 2020.

[32] 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, 2022.

[33] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations," *arXiv preprint arXiv:2301.11699*, 2023.

[34] M. Xu, J. Ma, and Y. Zhu, "Dual-diffusion: Dual conditional denoising diffusion probabilistic models for blind super-resolution reconstruction

in rsis," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.
[35] 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.
[36] Y. Zhu, K. Zhang, J. Liang, J. Cao, B. Wen, R. Timofte, and L. Van Gool, "Denoising diffusion models for plug-and-play image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1219–1229.
[37] C. Corneanu, R. Gadde, and A. M. Martinez, "Latentpaint: Image inpainting in latent space with diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4334–4343.
[38] R. Spetlik, D. Rozumnyi, and J. Matas, "Single-image deblurring, trajectory and shape recovery of fast moving objects with denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6857–6866.
[39] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 293–16 303.
[40] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 593–23 606, 2022.
[41] Z. Wang, Z. Zhang, X. Zhang, H. Zheng, M. Zhou, Y. Zhang, and Y. Wang, "Dr2: Diffusion-based robust degradation remover for blind face restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1704–1713.
[42] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, and B. Wen, "Shadowdiffusion: When degradation prior meets diffusion model for shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 049–14 058.
[43] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.
[44] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 683–25 696, 2022.
[45] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023.
[46] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1689–1697.
[47] B. Tu, X. Liao, Q. Li, C. Zhou, and A. Plaza, "Multi-resolution pyramid enhanced non-local feature extraction for hyperspectral classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 5865–5879, 2022.
[48] 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.
[49] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
[50] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *Advances in neural information processing systems*, vol. 32, pp. 64–69, 2019.
[51] J. Hai, R. Yang, Y. Yu, and S. Han, "Combining spatial and frequency information for image deblurring," *IEEE Signal Processing Letters*, vol. 29, pp. 1679–1683, 2022.
[52] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 561–10 570.
[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.
[54] 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.
[55] 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.
[56] 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.
[57] J. Li, K. Zheng, L. Gao, Z. Han, Z. Li, and J. Chanussot, "Enhanced deep image prior for unsupervised hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 20, 2023.
[58] 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.
[59] Z.-H. Cao, Y.-J. Liang, L.-J. Deng, and G. Vivone, "An efficient image fusion network exploiting unifying language and mask guidance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2025, Early Access.
[60] 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.
[61] Q. Zhu, M. Zhang, Y. Chen, and G. Zheng, "Spectral correlation-based fusion network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2024.
[62] 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.
[63] Y. Yang, H. Zhao, X. Huangfu, Z. Li, and P. Wang, "Vit-isrgan: A high-quality super-resolution reconstruction method for multi-spectral remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 3973 – 3988, 2025.