remote sensing

MDPI

Article

Global-Local-Structure Collaborative Approach for Cross-Domain Reference-Based Image Super-Resolution

Xiuxia Cai [1,†], Chenyang Diwu [2,†], Ting Fan [2], Wenjing Wang [2,*] and Jinglu He [2]

[1] School of Electronic Engineering, Xi'an University of Post and Telecommunications, Xi'an 710121, China; caixiuxia@xupt.edu.cn
[2] School of Communication and Information Engineering, Xi'an University of Post and Telecommunications, Xi'an 710121, China; dwcy@stu.xupt.edu.cn (C.D.); fting@stu.xupt.edu.cn (T.F.); jlhe20@xupt.edu.cn (J.H.)
* Correspondence: wjing@xupt.edu.cn
† These authors contributed equally to this work.

Highlights

What are the main findings?

- A degradation-aware diffusion-based super-resolution framework is proposed, which explicitly models complex and mixed degradations in remote sensing images through adaptive conditional priors.
- A dual-decoder recursive generation strategy effectively balances local detail recovery and global structural consistency, achieving superior robustness under both ideal and blind degradation settings.

What are the implications of the main finding?

- Explicit degradation modeling and structural regularization significantly improve the reliability of super-resolution for real remote sensing scenarios.
- The proposed framework provides a practical and extensible solution for high-fidelity remote sensing image enhancement in downstream Earth observation tasks.

Abstract

Remote sensing image super-resolution (RSISR) aims to reconstruct high-resolution images from low-resolution observations of remote sensing data to enhance the visual quality and usability of remote sensors. Real world RSISR is challenging owing to the diverse degradations like blur, noise, compression, and atmospheric distortions. We propose hierarchical multi-task super- resolution framework including degradation-aware modeling, dual-decoder reconstruction, and static regularization-guided generation. Speciffcally, the degradation-wise module adaptively characterizes multiple types of degradation and provides effective conditional priors for reconstruction. The dual-decoder platform incorporates both convolutional and Transformer branches to match local detail preservation as well as global structural consistency. Moreover, the static regularizing guided generation introduces prior constraints such as total variation and gradient consistency to improve robustness to varying degradation levels. Extensive experiments on two public remote sensing datasets show that our method achieves performance that is robust against varying degradation conditions.

Keywords: remote sensing; degradation-aware modeling; dual-decoder framework; static regularization

Academic Editor: Jon Atli Benediktsson

Received: 13 December 2025
Revised: 23 January 2026
Accepted: 30 January 2026
Published: 3 February 2026

1. Introduction

1.1. Background and Challenges

Remote sensing images play a critical role in acquiring large-scale spatial information and are widely applied in land resource surveying, urban planning, agricultural monitoring, ecological protection, military reconnaissance, and disaster assessment [1]. Beyond visual inspection, remote sensing imagery serves as a fundamental data source for numerous downstream intelligent interpretation tasks, including land-cover classification, semantic segmentation, change detection, and object detection. The performance of these high-level tasks is highly dependent on image quality, particularly spatial resolution and structural fidelity. High-resolution remote sensing images preserve rich textures and fine-grained geometric structures, which are essential for accurately recognizing small-scale ground objects and complex spatial patterns, such as narrow roads, building boundaries, vehicles, and farmland grids. However, in real-world acquisition scenarios, remote sensing images are inevitably affected by complex and mixed degradations, including limited sensor resolution, atmospheric turbulence, motion blur, cloud occlusion, and compression artifacts. These factors jointly lead to reduced spatial resolution, noise contamination, texture degradation, and blurred edges, significantly impairing both visual quality and the reliability of downstream analysis tasks. Consequently, remote sensing image super-resolution (RSISR) is not merely a low-level image enhancement problem, but a critical prerequisite for robust remote sensing interpretation systems. Early super-resolution approaches primarily relied on interpolation techniques (e.g., bicubic and Lanczos interpolation [2]) or sparse representation and dictionary learning methods [3,4]. Although computationally efficient, these methods are constrained by simplified priors and struggle to recover high-frequency textures and complex structures, often producing over-smoothed results. With the advancement of deep learning, convolutional neural networks (CNNs) [5,6] and Transformer-based architectures [7,8] have significantly improved reconstruction accuracy. Nevertheless, existing deep models still face challenges in RSISR, particularly in handling complex degradations and simultaneously preserving fine-grained local details and global structural consistency [9].

1.2. Related Work

Remote Sensing Image Super-Resolution. The objective of RSISR is to reconstruct high-resolution images from low-resolution observations to enhance spatial resolution and semantic recognition, providing reliable inputs for downstream tasks such as land-cover classification, object detection, and environmental monitoring [10,11]. Early interpolation-based approaches suffer from texture blurring and edge distortion [3], while sparse representation and dictionary learning methods improve detail recovery through sparsity constraints but exhibit limited adaptability under complex degradation conditions [4]. With the rise of deep learning, CNN-based methods such as EDSR [12] and RCAN [13] leverage deep residual learning and attention mechanisms to improve reconstruction accuracy. Transformer-based architectures, including SwinIR [7] and Uformer [8], further enhance global contextual modeling. However, most existing RSISR methods still rely on idealized degradation assumptions and struggle to achieve a balanced trade-off between local texture fidelity and global structural consistency in real-world scenarios [14].

Diffusion Models for Super-Resolution. Diffusion models have recently emerged as a powerful class of generative models due to their stable training process and strong distribution modeling capability [15–18]. In super-resolution tasks, diffusion-based approaches demonstrate superior perceptual quality, particularly in texture realism and detail diversity [19–21], and allow flexible conditional control through degradation parameters or reference images [22,23]. Recent studies have extended diffusion models to RSISR.

SinSR [24] introduces a single-step inference strategy to reduce computational cost, RefDiff [25] exploits reference images for structural detail transfer, and EDiffSR [26] incorporates structural constraints to improve reconstruction quality. Despite these advances, existing diffusion-based RSISR methods still suffer from simplified degradation assumptions, insufficient structural consistency in complex scenes, and instability under severe or mixed degradations [16,18,27,28].

Degradation Modeling and Structural Regularization. Degradation modeling plays a crucial role in RSISR by characterizing the transformation from high-resolution images to low-resolution observations. Most traditional approaches assume fixed degradation operators (e.g., bicubic downsampling), which are inadequate for remote sensing imagery affected by spatial blur, noise, compression artifacts, and atmospheric scattering [23,27]. Recent studies explore explicit degradation estimation [29] and implicit degradation representation learning [30], yet their generalization ability remains limited under heterogeneous and multi-modal degradation conditions. Regularization and structural priors are essential for stabilizing generative reconstruction. Techniques such as total variation regularization, gradient consistency constraints [28,31], and perceptual losses [19] help suppress artifacts and preserve edges. Recent evidence indicates that incorporating structural priors into diffusion models can significantly improve generation stability and reduce edge blurring [16,18,32]. However, static regularization strategies often lack adaptability across diffusion timesteps and degradation severities.

1.3. Motivation and Contributions

Motivated by the above challenges, we observe that existing degradation-aware diffusion-based super-resolution methods primarily treat degradation information as implicit conditions or noise-level embeddings, which are weakly coupled with structural modeling and often require multi-step stochastic sampling, making them particularly fragile in remote sensing scenarios with complex and heterogeneous degradations.

To address these limitations, we propose a global–local structure collaborative diffusion framework for remote sensing image super-resolution.

The main contributions of this work are summarized as follows:

- Unlike existing degradation-aware diffusion SR methods that rely on implicit or stochastic degradation conditioning, we propose an explicit degradation-aware modeling module that deterministically encodes multi-source and multi-scale degradation priors and injects them into the diffusion latent space.
- Different from prior diffusion-based RSISR frameworks that decouple local texture enhancement and global structural modeling, we design a dual-decoder global–local collaborative framework that tightly couples structural reconstruction with degradation-aware diffusion, enabling progressive and structurally consistent refinement.
- We introduce a static regularization guidance strategy in the diffusion latent space to stabilize structural preservation and improve perceptual quality.
- Extensive experiments on benchmark datasets demonstrate that the proposed method outperforms state-of-the-art approaches under both idealized and realistic degradation scenarios, showing strong robustness and generalization.

2. Methodology

2.1. Overall Framework

As shown in Figure 1, we develop a multi-module collaborative framework for remote sensing image super-resolution, specifically designed to address the structural blur and detail loss commonly observed in degraded images. Unlike existing degradation-aware diffusion-based SR methods that treat degradation information as implicit noise-

level embeddings and rely on multi-step stochastic sampling, our framework explicitly couples degradation modeling with structural reconstruction in a deterministic and efficient diffusion paradigm. The framework consists of three complementary modules. Degradation-Aware Modeling (DAM) module: In contrast to prior approaches that encode degradation implicitly, DAM explicitly models multi-source degradations and cross-scale blurs in remote sensing images. It extracts degradation vectors through lightweight CNNs with channel attention and injects them into the diffusion latent space as informative priors [27,30], enabling controlled and structure-aware generation. Dual-Decoder Design and Recursive Generation: This module integrates a local convolutional decoder and a global Transformer-based decoder to capture fine-grained textures and long-range semantic information. Through recursive refinement over diffusion timesteps and residual-domain corrections [28,31], the dual-decoder design progressively reconstructs images from coarse structures to fine details, ensuring semantic consistency and geometric accuracy. This design effectively alleviates structural instability under complex degradations. Static Regularization Guidance (SRG): SRG incorporates structural priors such as Total Variation and Gradient Consistency [19,32] into the diffusion latent space to guide edge preservation and texture enhancement. By adaptively adjusting regularization strength across diffusion stages, SRG stabilizes the generation process and maintains structural awareness, addressing limitations of previous methods in fine-detail control.

Overall, our framework explicitly integrates degradation modeling, global–local structural reconstruction, and prior-guided regularization to achieve high-quality, controllable, and structurally consistent super-resolution across multi-scene and multi-modal remote sensing images [20,21,23].

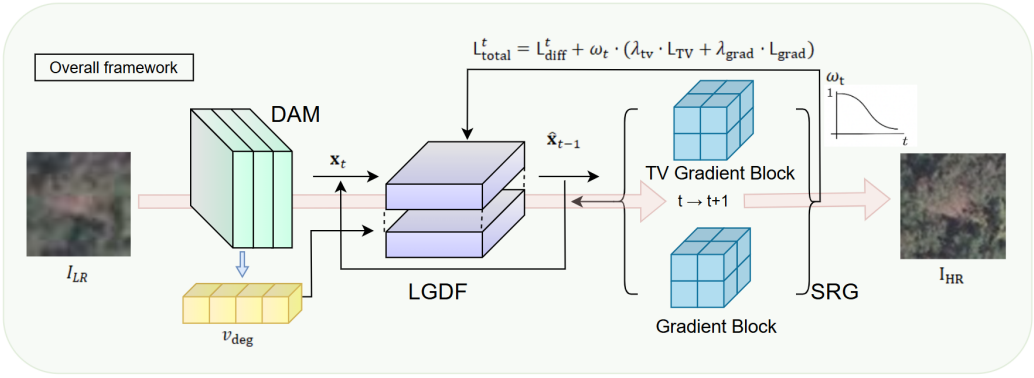

Figure 1. Overall Architecture of the Proposed Degradation-Aware Super-Resolution Framework with Dual-Decoder and Static Regularization-Guided Generation.

2.2. Degradation-Aware Modeling Module

Remote sensing images suffer from multiple-source and multi-scale degradations during acquisition, including spatial blur, compression noise, color distortion, and cloud occlusion [33–35]. These degradations are highly heterogeneous, spatially variant, and generally lack accurate annotations, making robust degradation modeling particularly challenging. Most existing degradation-aware diffusion-based super-resolution methods handle such degradations implicitly, for example by encoding them into noise-level embeddings or treating them as stochastic perturbations, which weakly couples degradation information with structural reconstruction and often leads to unstable results under complex real-world conditions. To address this limitation, we propose a Degradation-Aware Modeling (DAM) module, which explicitly extracts and represents degradation characteristics in a deterministic manner and serves as a structured guidance signal for the diffusion model, enabling more stable, accurate, and controllable super-resolution outcomes. Concep-

tually, DAM operates in three main steps: it first encodes multi-scale degradation features from the input image using a lightweight convolutional feature extractor [27,36], then aggregates and recalibrates these features through a channel attention mechanism into a compact degradation vector v_{deg} that encodes both the type and severity of degradative effects, as illustrated in Figure 2, and finally injects v_{deg} as an explicit prior into the diffusion model to steer the reconstruction process toward a more plausible and structurally consistent solution distribution [33,37]. Unlike existing degradation-aware diffusion SR methods that rely on implicit or stochastic degradation conditioning, DAM deterministically couples degradation modeling with structural reconstruction, improving stability, controllability, and cross-domain generalization under complex and heterogeneous degradation conditions.

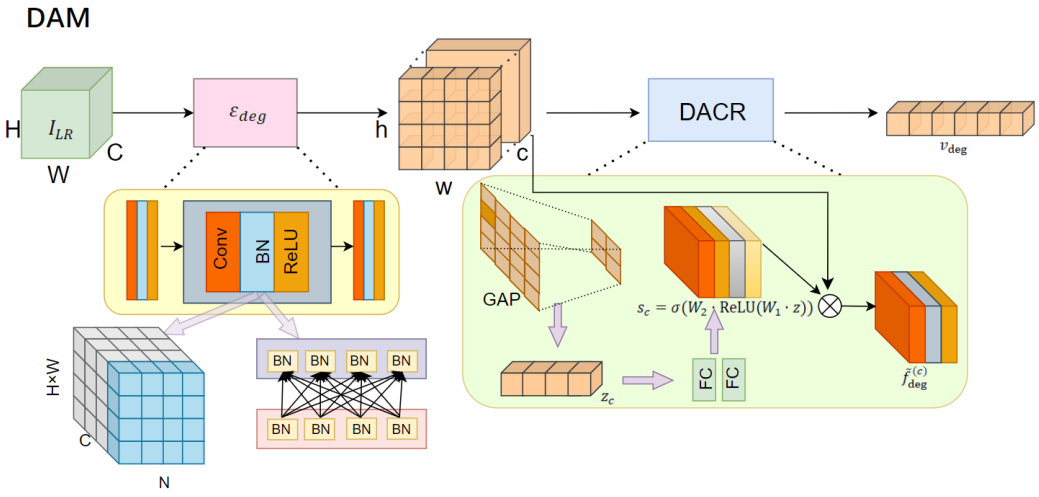

Figure 2. Degradation-Aware Modeling Module for Adaptive Representation of Complex Image Degradations.

The following subsections provide the detailed implementation of each step described above, including the lightweight feature extraction, channel-wise recalibration through the Degradation-Aware Channel Recalibration (DACR) mechanism, and conditional injection of the resulting degradation vector into the diffusion network. These implementations realize the deterministic guidance conceptualized in DAM, ensuring that the model focuses on degradation-sensitive regions and reconstructs structurally consistent high-resolution outputs.

2.2.1. Lightweight Feature Extractor

Given an input low-resolution remote sensing image $I_{\text{deg}} \in \mathbb{R}^{H\times W\times C}$, we create a lightweight convolutional neural network ε_{deg} which extracts degradation-related features. Our low-fidelity convolution layer consists of three light convolution layers with a 3×3 convolution kernel, batch normalization (BN), and ReLU activations:

$$f^{(i)} = \text{ReLU}\left(\text{BN}\left(\text{Conv}_{3\times 3}\left(f^{(i-1)}\right)\right)\right), \quad i = 1, 2, 3 \tag{1}$$

where $f^{(0)} = I_{\text{LR}}$. Finally, we obtain the degradation-aware feature map $f_{\text{deg}} \in \mathbb{R}^{h\times w\times c}$.

The BN layer is important for training and convergence. The idea is to normalize the mean and variance of each channel in a mini-batch to reduce internal covariate shift. Specifically, for each channel c, BN performs:

$$\hat{x}_{n,c,h,w} = \frac{x_{n,c,h,w} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \tag{2}$$

where μ_c and σ_c^2 denote the mean and variance of channel c, respectively, and ϵ is a small constant for numerical stability. Then, BN applies a learnable linear transformation with scaling and shifting parameters γ_c and β_c:

$$y_{n,c,h,w} = \gamma_c \hat{x}_{n,c,h,w} + \beta_c \tag{3}$$

This stabilizes the feature distributions across the network layers, removes vanishing or exploding gradients, and facilitates generalisation and training.

2.2.2. Degradation-Aware Channel Recalibration

To effectively capture the channel-wise response differences of multiple degradation types in remote sensing images, we propose a Degradation-Aware Channel Recalibration (DACR) mechanism. This module enhances the model's capability to represent degradation-sensitive regions through an adaptive channel weighting strategy. DACR is based on Squeeze-and-Excitation idea. It adjusts intermediate features along the channel dimension and emphasizes features closely associated with degradation patterns. Specifically, given the degradation-aware feature map $f_{\text{deg}} \in \mathbb{R}^{h \times w \times c}$, a global average pooling is first applied to each channel to extract its statistical descriptor:

$$z_c = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} f_{\text{deg}}^{(c)}(i, j) \tag{4}$$

The aggregated descriptor $z \in \mathbb{R}^c$ is then fed into a nonlinear transformation module consisting of two fully connected (FC) layers to model inter-channel dependencies and generate the attention weights $s_c \in \mathbb{R}^c$:

$$s_c = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \tag{5}$$

where $W_1 \in \mathbb{R}^{c/r \times c}$ and $W_2 \in \mathbb{R}^{c \times c/r}$ are the dimensionality-reduction and expansion matrices, r denotes the channel reduction ratio, and $\sigma(\cdot)$ is the Sigmoid activation function.

Subsequently, the generated attention weights are used to reweight the original feature map channels, producing the degradation-enhanced features:

$$\widetilde{f}_{\text{deg}}^{(c)} = s_c \cdot f_{\text{deg}}^{(c)} \tag{6}$$

The enhanced feature map is then flattened and projected into a low-dimensional degradation vector:

$$v_{\text{deg}} = W_f \cdot \text{vec}(\widetilde{f}_{\text{deg}}) + b_f \tag{7}$$

where $\text{vec}(\cdot)$ denotes the flattening operation, $W_f \in \mathbb{R}^{d \times hwc}$ and $b_f \in \mathbb{R}^d$ are the parameters of a fully connected layer.

The resulting degradation vector v_{deg} semantically encodes multi-type and multi-scale degradation information within the image, demonstrating strong discriminative and transferable capabilities. When added as an external prior, v_{deg} guides the diffusion model at each noise prediction step. It helps the model focus on important degraded regions. This improves the super-resolution quality and cross-domain performance under complex degradations.

2.2.3. Degradation-Aware Conditional Injection

The degradation vector v_{deg}, denoted as v hereafter, is injected into the diffusion network as a conditional guidance term, providing targeted degradation priors during each noise prediction step. In practice, this conditional injection enriches the diffusion

process with explicit knowledge of the degradation type and severity [33,35]. As a result, the model not only achieves higher accuracy in reconstructing images degraded by blur, noise, or low contrast, but also maintains robustness under heterogeneous degradation conditions [38,39]. More importantly, this mechanism demonstrates strong transferability in remote sensing applications. Since remote sensing images exhibit diverse spatial resolutions, sensor characteristics, and complex degradation patterns, super-resolution models trained solely on natural images often struggle to generalize. By incorporating the degradation-aware conditional injection, our model effectively leverages intrinsic degradation cues, bridging the gap between natural image pretraining and remote sensing image applications [33,40]. In this way, the proposed approach enables cross-modal and cross-task generalization, which is crucial for practical deployment in real-world remote sensing scenarios. In this way, the explicit degradation vector realizes the deterministic guidance of DAM, closing the loop between conceptual modeling and practical implementation.

2.3. Dual-Decoder Design and Recursive Generation

In single-scale conditional diffusion modeling, we propose structure-aware dual-decoder design and cross-time recursive generation by progressive reconstruction from coarse structures to fine details [41]. The model is capable of capturing high-frequency structures and edges in remote sensing images. It works well for reconstructing complex textures and handling various degradations. Specifically, the proposed module integrates multi-stage recursion along the temporal dimension (Time-wise Recursion) with layer-wise residual refinement in the residual domain, thereby achieving progressive optimization from structure to texture [42]. Although the model employs feature-level recursion for progressive refinement, the final high-resolution output is generated in a single forward pass, i.e., our framework operates as a single-step diffusion model.

2.3.1. Dual-Decoder Design

In classical diffusion models, the decoder is usually implemented as a single architecture such as U-Net, which iteratively denoises the Gaussian corrupted image x_T into a clean target. However, single-decoder structures often struggle with complex degradations in remote sensing images, such as non-uniform blur, occluded structures, and multi-scale texture damage. This can cause missing local details and inconsistent semantics. To solve this problem, we propose a Local-Global Decoding Framework (LGDF), shown in Figure 3. It is integrated into the single-scale conditional diffusion process. The goal is to jointly optimize global semantic modeling and local detail restoration. It consists of two parallel decoding branches: a local decoder D_L responsible for fine-grained structure reconstruction, and a global decoder D_G responsible for long-range semantic consistency.

(1) Decoding Output Formulation

The general reverse diffusion objective is defined as:

$$\hat{x}_{t-1} = x_t - \epsilon_\theta(x_t, t \mid v) \tag{8}$$

where $\epsilon_\theta(\cdot)$ is the noise predictor and $\mathbf{v}$ is the degradation condition vector. In LGDF the noise estimation term is obtained by fusing the outputs of the two decoders:

$$\epsilon_\theta(x_t, t \mid v) = \alpha \cdot \mathsf{D}_L(x_t, t, v) + (1 - \alpha) \cdot \mathsf{D}_G(x_t, t, v) \tag{9}$$

where $\alpha \in [0, 1]$ is a learnable fusion coefficient that adaptively balances the contributions of the two branches.

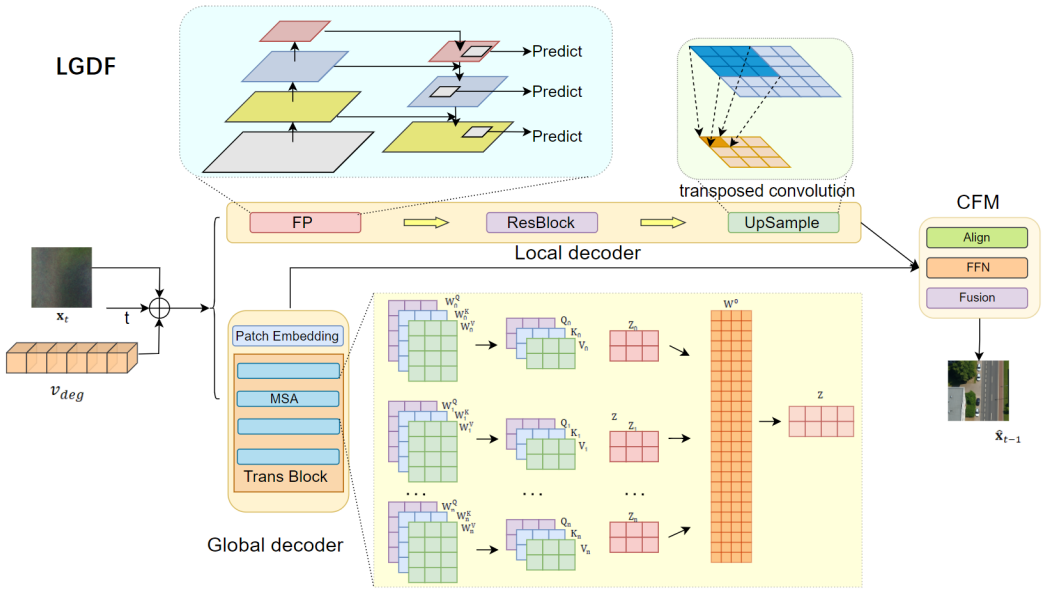

Figure 3. Dual-Decoder Architecture for Joint Structural Fidelity Preservation and Fine-Grained Texture Enhancement in Remote Sensing Image Super-Resolution.

(2) Local Decoder

Local branch D_L adopts a U-Net to enhance local detail reconstruction. Features pyramids, residual connections, and upsampling modules are fused across-scale. The overall state of the local decoding is:

$$D_L(x_t, t, v) = \text{UpSample} \circ \text{ResBlock} \circ \text{FP}(x_t \oplus \phi(t) \oplus v) \tag{10}$$

where $\phi(t)$ denotes the temporal embedding vector, and $\oplus$ represents channel-wise concatenation. ResBlock introduces local residual learning to improve the representation of edges and textures [43]. The feature pyramid $\text{FP}(\cdot)$ is constructed as:

$$D_L(x_t, t, v) = \text{UpSample} \circ \text{ResBlock} \circ \text{FP}(x_t \oplus \phi(t) \oplus v) \tag{11}$$

where Conv_k denotes the convolution operator at scale k for processing hierarchical features, and K is the total number of scales [44]. The upsampling using transposed convolution is used to update the features to high resolution.

The final result is high-frequency details and geometric consistency which supports local details recovery for the entire decoder and leads to local sharpness and texture quality in degraded areas. The same works for global branch. So, the decoder produces consistent semantics and better quality images.

(3) Global Decoder

The global branch D_G is based on multiple Transformer blocks to capture long-range dependencies and global consistency [45]. The branch first compiles the input features into a patch by Patch Embedding and then applies a Multi-Head Self-Attention (MSA) to model global feature.The overall process is formulated as:

$$D_G(x_t, t, v) = \text{MSA} \circ \text{PatchEmbed}(x_t \oplus \phi(t) \oplus v) \tag{12}$$

where $\phi(t)$ denotes the temporal embedding vector, and $\oplus$ represents channel-wise concatenation.

In the global decoder, the main operation is MSA. In this case, all input features are first projected to the token sequence **z** with Patch Embedding. For each attention head, different linear projection matrices W^Q, W^K, and W^V generate the query (Q), key (K), and value (V) representations, each head computes a weighted output to capture global dependencies across subspaces. The multi-head outputs are concatenated and linearly projected using W^O to produce the fused contextual representation z.

Each Transformer block integrates an MSA module with a residual connection, defined as:

$$z' = \mathrm{MSA}(z) + z \tag{13}$$

where z denotes the input token sequence.

This global branch focuses on modeling structural and semantic consistency, effectively addressing non-local degradation issues such as large-area occlusion, geometric distortion, and artifacts. By providing macro-level semantic constraints, it complements the local decoder and contributes to achieving both global coherence and perceptually faithful reconstruction.

(4) Complementary Fusion Module

In order to combine the structure features of convolutional local decoder D_L with semantic features of Transformer global decoder D_G, we design a Complementary Fusion Module (CFM). CFM consists of three types of features alignment, feature fusion, and nonlinear interaction modeling.

Feature alignment stage. Since the outputs of D_L and D_G differ in channel dimension and semantic distribution, we first apply linear mappings to align feature spaces:

$$F_L^{\text{align}} = W_L F_L + b_L, \quad F_G^{\text{align}} = W_G F_G + b_G \tag{14}$$

Feature fusion stage. The aligned local and global features are concatenated:

$$F_{\text{fusion}} = \mathrm{Concat}(F_L^{\text{align}}, F_G^{\text{align}}) \tag{15}$$

Nonlinear interaction stage. A lightweight feed-forward network (FFN) is introduced to enhance semantic interactions across channels:

$$\hat{F}_{\text{fusion}} = F_{\text{fusion}} + \mathrm{FFN}(F_{\text{fusion}}), \quad \mathrm{FFN}(x) = W_2 \cdot \sigma(W_1 x + b_1) + b_2 \tag{16}$$

where $\sigma(\cdot)$ is activation. Finally, we use the last fused representation $\hat{F}_{\text{fusion}}$ for subsequent reconstruction, which gives much richer expressivity for multi-level structures and semantic details.

(5) Inference Process and Computational Characteristics

During inference, a low-resolution image $\mathbf{x}_0$ and its degradation vector **v** are fed into the model in a single forward pass. The local decoder D_L extracts fine-grained structural details, while the global decoder D_G captures long-range semantic context. The outputs of both branches are fused through the CFM to produce the final high-resolution reconstruction $\hat{\mathbf{x}}$. No iterative multi-step diffusion sampling is performed; the entire reconstruction is achieved deterministically in one forward pass.

Compared to conventional multi-step diffusion models, the single-step forward design substantially reduces inference time and memory consumption. The dual-decoder and CFM modules provide rich feature representation and maintain structural fidelity without the need for multiple denoising iterations, resulting in both efficient computation and stable reconstruction quality for complex remote sensing images.

2.3.2. Recursive Generation Mechanism

We also propose a recursive generation function that iteratively improves features from two perspectives. It should be noted that both time-wise recursion and residual correction recursion operate at the feature level for progressive refinement, rather than performing multi-step diffusion sampling; the final high-resolution output is still produced in a single forward pass. First, time-wise recursion performs progressive denoising of several timesteps, while residual correction recursion optimizes structural residuals to better capture edges and textures. The two mechanisms provide complementary models: time-wise recursion improves global stability and semantic coherence, and residual recursion enhances local detail quality [22]. Both of them lead to high perceptual quality and local readability [18].

(1) Time-wise Recursion

The fundamental idea of diffusion models is to progressively generate a clean image from pure noise through a reverse denoising process. Based on this result, we introduce a cross-timestep recursive mechanism where the result of a reconstruction at the current timestep is fed back as a prior for a next timesteps. The structure is shown in Figure 4. This enables a coarse-to-fine progressive generation process. The mechanism can be formally expressed as:

$$\hat{x}_{t-1} = \mathcal{R}_{t-1}(x_t, v, \hat{x}_t) \tag{17}$$

where $\hat{\mathbf{x}}_t$ is the estimated image at timesteps t, $\mathbf{v}$ is the degradation vector, and $\mathcal{R}_{t-1}(\cdot)$ is the recursive restore function based on the current estimate and the degradation prior between timestep t and $t-1$. We define the recursive function as follows:

$$\mathcal{R}_{t-1}(x_t, v, \hat{x}_t) = f_\theta(x_t, \hat{x}_t, \phi(v)) \tag{18}$$

where f_θ denotes the recursive restoration network, and $\phi(\mathbf{v})$ is the embedded representation of the degradation vector. Roughly, if we can update this formulation, a high quality image can be modelled in steps as:

$$\hat{x}_0 = \mathcal{R}_0\Big(\mathcal{R}_1(\ldots \mathcal{R}_{T-1}(x_T))\Big) \tag{19}$$

Step-by-step recursion, starting from $\mathbf{x}_T \sim \mathcal{N}(0, I)$, reflects the time evolution behavior of diffusion models. In a multi-step network, such recursive dynamics are shown to yield more accurate detail recovery and uncertainty model.

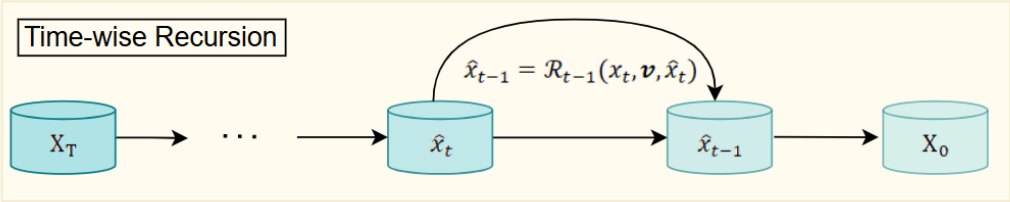

Figure 4. The time-wise recursion progressively refines image generation across multiple timesteps, where each estimated result $\hat{x}_t$ serves as a prior for the next step to enhance temporal consistency and semantic stability.

(2) Residual Correction Recursion

Although the primary diffusion-based reconstruction path can generate high-quality preliminary images, reconstruction deficiencies still exist in complex object boundaries and weak texture regions of remote sensing imagery. Therefore, as shown in the Figure 5,

we introduce a structural residual correction mechanism. Specifically, based on the reconstructed image from the main diffusion path, a lightweight residual prediction network $\mathcal{C}(\cdot)$ is employed to model and compensate for the residual components:

$$\hat{x}_0^{\text{final}} = \hat{x}_0 + \mathcal{C}(\hat{x}_0, v) \tag{20}$$

Here, $\mathcal{C}(\cdot)$ consists of a shallow convolutional feature extraction module and a channel attention enhancement module (CA), which jointly estimate the structural residual information using both the reconstructed image and the degradation vector [46]. The shallow convolutional feature extraction part focuses on learning local structural features and texture variations within the residual, implemented through two consecutive 3×3 convolutional layers with ReLU activations:

$$F_{conv} = \text{ReLU}(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{3\times3}(F_{in})))) \tag{21}$$

The channel attention enhancement module introduces a lightweight cross-channel interaction design to replace conventional fully connected attention structures. This mechanism applies a 1D convolution to the channel descriptor vector from global average pooling. It captures local relationships between channels without changing the feature dimension. This improves feature discrimination while keeping computation efficient. The process is formulated as follows:

$$z = \text{GAP}(F_{conv}) \in \mathbb{R}^{1\times1\times C} \tag{22}$$

$$w = \sigma(\text{Conv1D}(z)) \tag{23}$$

$$F_{att} = F_{conv} \odot w \tag{24}$$

where $\odot$ denotes channel-wise multiplication. A final 3×3 convolution is then used to transform the weighted features into a structural residual map:

$$R = \text{Conv}_{3\times3}(F_{att}) \tag{25}$$

which is added to the current reconstruction result to complete one correction step:

$$\hat{x}_0^{(k+1)} = \hat{x}_0^{(k)} + R \tag{26}$$

The recursive correction process can be repeated several times, forming:

$$\hat{x}_0^{(k+1)} = \hat{x}_0^{(k)} + \mathcal{C}(\hat{x}_0^{(k)}, v), \quad k = 0, 1, \ldots, K-1 \tag{27}$$

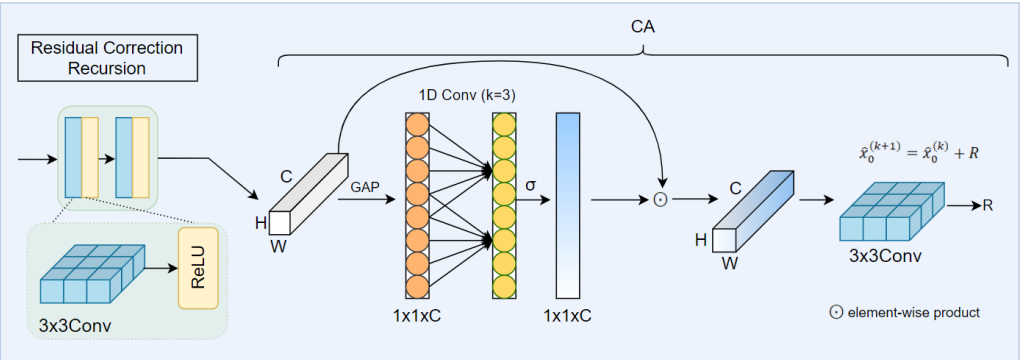

Figure 5. The residual correction recursion refines the reconstruction by predicting structural residuals through convolutional and channel attention modules, enabling iterative enhancement of fine details and edge structures.

This mechanism introduces an adaptable optimizer at the network level which progressively optimizes the reconstruction with recursive residual compensation. Compared to static post-processing or one-shot prediction, the recursive correction mechanism is more flexible and sensitive to structure, it better reproduces geometric consistency and sharpness of the details in remote sensing image reconstruction.

2.4. Static Regularization Guidance

Edge structures and texture details play a major role in remote sensing image super resolution tasks for perceptual quality and semantic discrimination. Although diffusion models can fit data distributions well and generate diverse results, they often blur boundaries and lose details. This is especially true in transition areas between land cover types, such as building-to-ground or water-to-vegetation. Introducing static structural priors not only enhances the structural perception ability of the generated images but also significantly improves texture representation and edge sharpness in natural regions such as water, forest, and roads. Prior studies (e.g., DPS [47], IDDPM [22]) have demonstrated that incorporating prior knowledge into the diffusion process can improve both convergence stability and controllability of generation [22].

Based on this idea, we add a static structural prior regularization. It explicitly constrains the structure of generated images. This improves the model's sensitivity and accuracy for edges and textures. During each reconstruction stage $\hat{I}_t^{SR}$ of the diffusion process, the static regularization terms are integrated into the optimization objective to balance structural preservation and detail reconstruction at different time steps. Specifically, two types of regularization are included: Total Variation (TV) regularization and Gradient Consistency Loss.

2.4.1. Overall Loss Function

On top of the primary diffusion reconstruction loss, we incorporate structural regularization to construct the following joint optimization objective:

$$\mathcal{L}_{\text{total}}^t = \mathcal{L}_{\text{diff}}^t + \omega_t \cdot \left(\lambda_{tv} \cdot \mathcal{L}_{TV} + \lambda_{grad} \cdot \mathcal{L}_{grad} \right), \tag{28}$$

where $\mathcal{L}_{\text{diff}}^t$ denotes the reconstruction loss in the diffusion process, including L1 loss and perceptual loss; $\mathcal{L}_{TV}$ is the total variation regularization term that suppresses local artifacts and oscillations [48]; and $\mathcal{L}_{grad}$ is the gradient consistency loss enforcing similarity in the gradient domain between generated images and high-resolution references [49]. The weighting coefficients λ_{tv} and λ_{grad} are introduced to balance artifact suppression and structural fidelity. In remote sensing imagery, excessive TV regularization may lead to over-smoothing of fine textures and small-scale structures, while gradient consistency plays a more critical role in preserving edges and geometric boundaries. Therefore, λ_{tv} is set to a relatively small value (0.1) to provide mild smoothness constraints, whereas λ_{grad} is assigned a larger weight (1.0) to emphasize structural preservation. These values are empirically determined and found to provide stable and robust performance across different datasets and degradation settings.

To adaptively adjust the influence of structural regularization across different diffusion time steps, we design a dynamic weighting factor ω_t. It balances the strength of structural guidance between the early and late stages of the diffusion process. Specifically, $\omega_t \in [0, 1]$ decreases over time. First of all, it imposes strong structural constraints for robust coarse reconstruction. Then the constraints are weaker in order to maintain fine details and avoid smoothing. We adopt a cosine scheduling strategy, formally defined as:

$$\omega_t = \cos\left(\frac{\pi t}{2T}\right), \tag{29}$$

where T is the number of diffusion steps and $t \in [0, T]$ is the total number of steps that are made. With this scheduling, regularization strength is smoothly decay, and the structure stability is complemented with detail, yielding high quality and visual realism.

2.4.2. Total Variation Regularization

Total variation is a classical image smoothing and denoising technique to enhance the structural sharpness at edges. It is defined as:

$$\mathcal{L}_{TV} = \sum_{i,j}\left(\left|\nabla_x \hat{I}^{SR}_{i,j}\right| + \left|\nabla_y \hat{I}^{SR}_{i,j}\right|\right), \tag{30}$$

where $\hat{I}^{SR}$ is the current image and ∇_x, ∇_y is a first order gradient operator in horizontal and vertical directions. It is calculated on the final output $\hat{I}^{SR}_0$ of the diffusion and added as a regularization to the main loss. It guides the generation in structural consistency and edge clarity. It minimizes high-frequency oscillatory artifacts and avoids smoothing.

2.4.3. Gradient Consistency Loss

In order to further facilitate structural alignment between generated images and high resolution ground truth, we present a gradient consistency loss:

$$\mathcal{L}_{grad} = \left\|\nabla \hat{I}^{SR} - \nabla I^{HR}\right\|_1, \tag{31}$$

where I^{HR} denotes the high-resolution reference image and ∇ represents the Sobel gradient operator. This loss maintains consistency both in edge direction and intensity, especially in transition regions (e.g., blurred borders, blurred textures). This loss has the effect of making the fine details more consistent.

3. Results

3.1. Datasets and Evaluation Metrics

UCMerced LandUse Dataset

The UCMerced LandUse dataset [50] contains 21 classes of typical land-use categories, including residential areas, farmland, forest, rivers, etc., with a total of 2100 aerial remote sensing images. Each image has a spatial resolution of 256 × 256 pixels, exhibiting rich texture and structural features across diverse natural and man-made scenes. The dataset can be used to evaluate the quality of restoration of details. In our experiments, 1800 images are used for training, 300 for validation, and the remaining 300 for testing.

AID Dataset

The AID dataset [51] contains 30 representative remote sensing scene categories containing 10,000 aerial images.The image resolution ranges from 600 × 600 to 1000 × 1000 pixels. Due to the large number of scenes and large scale variability, this information may be used to evaluate model robustness in difficult situations. To unify input resolutions, we crop the original images into 256 × 256 patches, resulting in 8000 training samples, 1000 validation samples, and 1000 test samples.

Evaluation Metrics

The reconstructed results are compared based on accuracy (PSNR, SSIM), perceptual quality (LPIPS), error magnitude (RMSE) [52], and fidelity (VIF) [53]. PSNR and SSIM measure the pixel-wise accuracy and structural similarity between the reconstructed images and the ground truth, respectively. LPIPS is employed as a perceptual metric that

evaluates similarity in deep feature space, which correlates well with human visual perception and is particularly relevant for assessing texture realism and structural consistency in remote sensing images. RMSE quantifies the pixel-wise error magnitude, while VIF measures the amount of visual information preserved from the reference image, making it suitable for evaluating detail and structural fidelity under complex degradations commonly encountered in remote sensing scenarios. Together, these metrics provide a comprehensive evaluation of reconstruction accuracy, perceptual quality, and information fidelity for remote sensing image super-resolution tasks.

3.2. Comparison with Existing Methods

Experimental Setup

To comprehensively evaluate the performance of the proposed method, we selected several representative methods for comparison in remote sensing image super-resolution tasks. The diffusion-based methods include SinSR, EDiffSR, and RefDiff, while the non-diffusion-based methods include EDSR [12], RCAN [13], SwinIR [7], and Uformer [8]. These methods cover representative and widely adopted state-of-the-art approaches across CNN-based, Transformer-based, and diffusion-based paradigms, which have been extensively used as strong baselines in recent super-resolution studies. Although several super-resolution methods have been proposed in the past two years, some of them are not publicly available or are evaluated under different experimental settings, making fair and reproducible comparisons challenging. Therefore, the selected baselines are considered sufficient to provide a comprehensive and fair evaluation of the proposed method, while recent SOTA methods are additionally discussed in the revised manuscript for better contextualization. The experiments cover three upscaling factors ($\times 2$, $\times 3$, $\times 4$) and two types of degradation scenarios. The first is Bicubic degradation, where low-resolution inputs are generated solely via bicubic downsampling for benchmark evaluation. The second is Blind degradation, in which, in addition to downsampling, multiple degradation factors such as Gaussian blur, additive noise, and optional JPEG compression are applied. All degradation parameters are randomly sampled within reasonable ranges to better simulate the degradation process of real remote sensing images.

All methods are evaluated under the same data splits, training epochs, and hardware environment, strictly following official implementations or publicly released training configurations. Training uses the Adam optimizer with an initial learning rate of 2×10^{-4}, decayed via a cosine annealing schedule. The batch size is 16, and all models are trained for 500 epochs on an NVIDIA RTX 4090 GPU (CUDA 12.2). The UCMerced LandUse and AID datasets are tested. PSNR, SSIM, LPIPS, RMSE, and VIF are used to evaluate pixel accuracy, perceptual quality, and information fidelity of the reconstructed images.

Fairness of Comparisons

To ensure fair evaluation, all baseline methods are trained and tested on the same datasets (UCMerced and AID) with identical data splits, training schedules, batch sizes, and hardware environments. Official or publicly released implementations and recommended hyperparameters are strictly followed. We note that some baselines, such as EDSR and RCAN, are originally optimized for ideal bicubic degradation, while our method also addresses blind and mixed degradations. Despite these differences in prior assumptions and task settings, the reported results (Tables 1 and 2) provide a meaningful and fair comparison, and any remaining discrepancies are acknowledged in the analysis.

Table 1. Quantitative comparison of different methods under bicubic degradation. Metrics: PSNR↑/SSIM↑/LPIPS↓/RMSE↓/VIF↑ (↑ higher is better; ↓ lower is better). Red text indicates the best, blue text the second-best, and purple background highlights important data.

Scale	Method	UCMerced	AID
×2	EDSR	32.14/0.918/0.112/0.025/2.0	31.02/0.912/0.125/0.028/2.1
	RCAN	32.48/0.923/0.108/0.023/2.1	31.30/0.917/0.120/0.026/2.2
	SwinIR	32.60/0.925/0.106/0.022/1.9	31.40/0.919/0.118/0.025/2.0
	Uformer	32.35/0.922/0.109/0.024/2.0	31.25/0.916/0.121/0.027/2.1
	SinSR	32.55/0.924/0.107/0.023/2.0	31.38/0.918/0.119/0.026/2.0
	RefDiff	32.50/0.923/0.108/0.023/2.0	31.35/0.917/0.120/0.026/2.1
	EDiffSR	32.58/0.924/0.107/0.022/1.9	31.39/0.918/0.119/0.025/2.0
	Ours	33.12/0.932/0.098/0.020/1.8	32.01/0.926/0.107/0.022/1.8
×3	EDSR	30.50/0.870/0.140/0.033/2.1	29.45/0.858/0.155/0.035/2.2
	RCAN	30.80/0.875/0.135/0.031/2.2	29.70/0.862/0.150/0.033/2.3
	SwinIR	31.00/0.878/0.132/0.030/2.0	29.85/0.865/0.148/0.032/2.1
	Uformer	30.75/0.873/0.136/0.032/2.1	29.65/0.860/0.151/0.034/2.2
	SinSR	30.95/0.876/0.134/0.031/2.0	29.80/0.863/0.149/0.033/2.1
	RefDiff	30.90/0.875/0.135/0.031/2.1	29.75/0.862/0.150/0.033/2.1
	EDiffSR	30.97/0.876/0.134/0.030/2.0	29.82/0.863/0.149/0.032/2.1
	Ours	31.55/0.888/0.125/0.028/1.9	30.20/0.875/0.138/0.030/1.9
×4	EDSR	28.90/0.820/0.170/0.038/2.2	27.80/0.805/0.185/0.040/2.3
	RCAN	29.30/0.825/0.165/0.036/2.3	28.20/0.810/0.180/0.038/2.4
	SwinIR	29.45/0.828/0.162/0.035/2.1	28.35/0.813/0.178/0.037/2.2
	Uformer	29.20/0.823/0.166/0.036/2.2	28.15/0.808/0.181/0.038/2.3
	SinSR	29.40/0.826/0.163/0.035/2.1	28.33/0.811/0.179/0.037/2.2
	RefDiff	29.35/0.825/0.164/0.035/2.2	28.30/0.810/0.180/0.037/2.3
	EDiffSR	29.42/0.826/0.163/0.035/2.1	28.34/0.811/0.179/0.037/2.2
	Ours	30.10/0.838/0.150/0.032/1.9	29.05/0.823/0.163/0.034/1.9

Table 2. Quantitative comparison of different methods under bicubic degradation. Metrics: PSNR↑/SSIM↑/LPIPS↓/RMSE↓/VIF↑ (↑ higher is better; ↓ lower is better). Red text indicates the best, blue text the second-best, and purple background highlights important data.

Scale	Method	UCMerced	AID
×2	EDSR	31.20/0.905/0.125/0.028/2.1	30.10/0.898/0.138/0.030/2.2
	RCAN	31.55/0.910/0.120/0.026/2.2	30.40/0.903/0.133/0.028/2.3
	SwinIR	31.70/0.913/0.118/0.025/2.0	30.55/0.905/0.131/0.027/2.1
	Uformer	31.45/0.911/0.121/0.027/2.1	30.35/0.903/0.134/0.029/2.2
	SinSR	31.65/0.912/0.119/0.026/2.1	30.50/0.905/0.132/0.028/2.1
	RefDiff	31.60/0.911/0.120/0.026/2.1	30.45/0.904/0.133/0.028/2.2
	EDiffSR	31.68/0.912/0.119/0.025/2.0	30.52/0.905/0.132/0.027/2.1
	Ours	32.70/0.925/0.105/0.022/1.9	31.55/0.916/0.118/0.024/1.9
×3	EDSR	29.45/0.885/0.142/0.033/2.2	28.40/0.872/0.157/0.035/2.3
	RCAN	29.80/0.890/0.138/0.031/2.3	28.75/0.877/0.152/0.033/2.4
	SwinIR	29.95/0.892/0.135/0.030/2.1	28.90/0.880/0.149/0.032/2.2
	Uformer	29.70/0.889/0.139/0.032/2.2	28.70/0.876/0.153/0.034/2.3
	SinSR	29.90/0.891/0.136/0.031/2.1	28.85/0.879/0.150/0.033/2.2
	RefDiff	29.85/0.890/0.137/0.031/2.2	28.80/0.878/0.151/0.033/2.2
	EDiffSR	29.92/0.891/0.136/0.030/2.1	28.87/0.879/0.150/0.032/2.2
	Ours	30.95/0.905/0.121/0.028/1.9	29.90/0.893/0.134/0.030/1.9
×4	EDSR	27.90/0.860/0.165/0.038/2.3	26.85/0.848/0.180/0.040/2.4
	RCAN	28.35/0.868/0.160/0.036/2.4	27.30/0.855/0.175/0.038/2.5
	SwinIR	28.50/0.871/0.157/0.035/2.2	27.45/0.858/0.172/0.037/2.3
	Uformer	28.25/0.868/0.161/0.036/2.3	27.25/0.855/0.174/0.038/2.4
	SinSR	28.45/0.870/0.158/0.035/2.2	27.40/0.857/0.173/0.037/2.3
	RefDiff	28.40/0.869/0.159/0.035/2.3	27.35/0.856/0.174/0.037/2.4
	EDiffSR	28.48/0.870/0.158/0.035/2.2	27.42/0.857/0.173/0.037/2.3
	Ours	29.50/0.882/0.142/0.032/1.9	28.45/0.871/0.155/0.034/1.9

Quantitative Results Analysis

The overall quantitative results are summarized in Tables 1 and 2, evaluated on the UCMerced and AID datasets under both ideal and non-ideal degradation settings. The employed metrics jointly reflect pixel-level fidelity (PSNR, SSIM), perceptual quality (LPIPS), reconstruction error (RMSE), and information preservation capability (VIF), which are particularly important for remote sensing image interpretation.

Ideal degradation: Under the ideal bicubic degradation setting, the proposed method achieves performance comparable to state-of-the-art super-resolution approaches such as RCAN [13] and SwinIR [7] in terms of PSNR and SSIM, especially for the $\times 2$ upscaling task. This indicates that the proposed diffusion-based framework does not compromise pixel-level reconstruction accuracy. This behavior can be attributed to the dual-decoder architecture, which explicitly separates structural reconstruction from fine-detail refinement, together with the static regularization guidance that constrains the diffusion process and suppresses excessive smoothing.In terms of perceptual quality, the proposed method consistently outperforms non-diffusion methods and achieves comparable LPIPS scores to diffusion-based approaches such as SinSR [24] and RefDiff [25], demonstrating its ability to maintain perceptual realism even under mild degradations. Meanwhile, the lower RMSE and higher VIF values indicate that the proposed method preserves more informative content and structural details, which is critical for remote sensing scenes containing man-made objects and fine textures.

Non-ideal degradation: Under non-ideal and blind degradation conditions, the performance gap between different methods becomes more pronounced. Non-diffusion-based approaches (EDSR [12], RCAN [13], SwinIR [7]) suffer from significant degradation in PSNR and SSIM (typically around 1–2 dB), accompanied by inferior LPIPS and VIF scores, reflecting their limited robustness to complex degradation combinations. Although diffusion-based methods retain advantages in perceptual metrics, some methods (e.g., RefDiff [25]) exhibit suboptimal RMSE and VIF, suggesting instability in structural and information preservation under strong noise and mixed degradations. In contrast, the proposed method demonstrates consistently superior performance across all metrics. Specifically, it achieves approximately +1.3 dB PSNR and +0.018 SSIM improvements over RCAN [13] for $\times 2$–$\times 4$ tasks, along with a 10–20% reduction in LPIPS, lower RMSE, and notably higher VIF. These improvements can be directly attributed to the degradation-aware modeling module (DAM), which explicitly encodes degradation characteristics and enables adaptive reconstruction strategies for different degradation types. By guiding the diffusion process with degradation-aware priors and regularization constraints, the proposed framework effectively mitigates error accumulation and hallucinated artifacts, resulting in better preservation of structural and edge information. The advantages are especially evident under high upscaling factors ($\times 4$) and strong noise degradation, highlighting the robustness of the proposed design.

Qualitative Results and Analysis

Figures 6 and 7 provide qualitative comparisons to further validate the effectiveness of the proposed method. Figure 6 presents reconstruction results under ideal bicubic degradation on the UCMerced dataset, while Figure 7 shows results under blind degradation conditions on the AID dataset. Under ideal degradation, most methods can recover the global structure of scenes; however, CNN- and Transformer-based baselines such as RCAN [13], SwinIR [7], and Uformer [8] tend to produce over-smoothed textures and blurred edges. In contrast, the proposed method reconstructs finer textures and sharper boundaries, such as building edges, tennis court markings, and vehicle contours. These visual improvements are consistent with the observed gains in PSNR, SSIM, and LPIPS, indicating that the proposed architecture enhances perceptual quality without sacrificing

pixel accuracy. Under the more challenging non-ideal degradation setting, the superiority of the proposed approach becomes more evident. Some diffusion-based methods (e.g., RefDiff [25], EDiffSR [26]) generate structural inconsistencies or artifacts when facing mixed blur and noise degradations. Benefiting from the degradation-aware modeling and regularization-guided diffusion generation, the proposed method maintains geometric consistency and reliable detail reconstruction across complex scenes, including urban areas, vegetation regions, and dense building layouts. Overall, the qualitative results provide intuitive evidence for the quantitative improvements, demonstrating strong robustness and cross-dataset generalization capability.

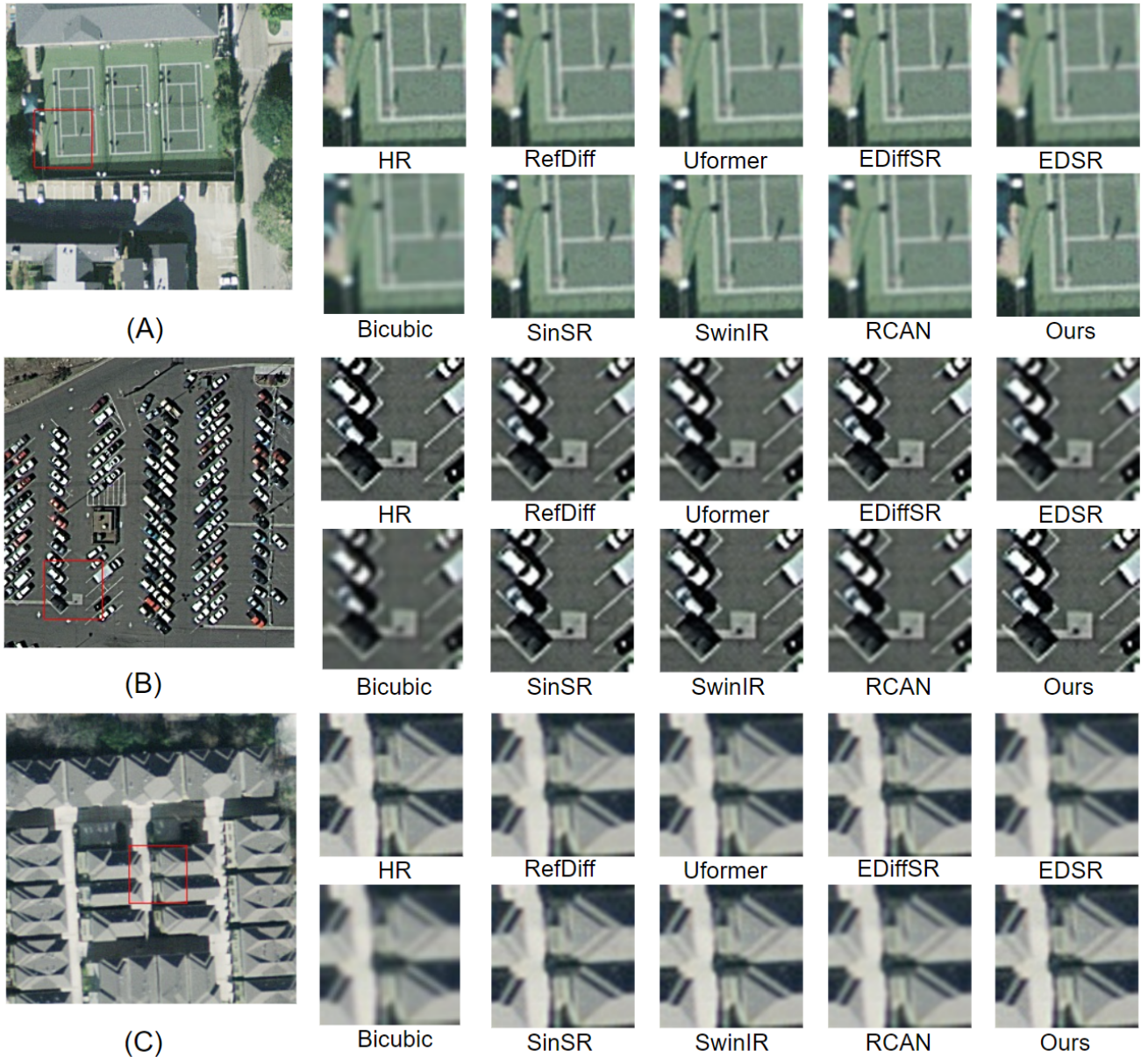

Figure 6. Qualitative comparison of different methods at ×3 scale under bicubic degradation on the UCMerced dataset. Three groups of images are shown: (**A**) first group, (**B**) second group, (**C**) third group. For each group, results from different methods are presented, and the proposed method reconstructs sharper edges and finer textures compared with existing CNN- and Transformer-based baselines. The red square highlights the region of interest in each image for comparison.

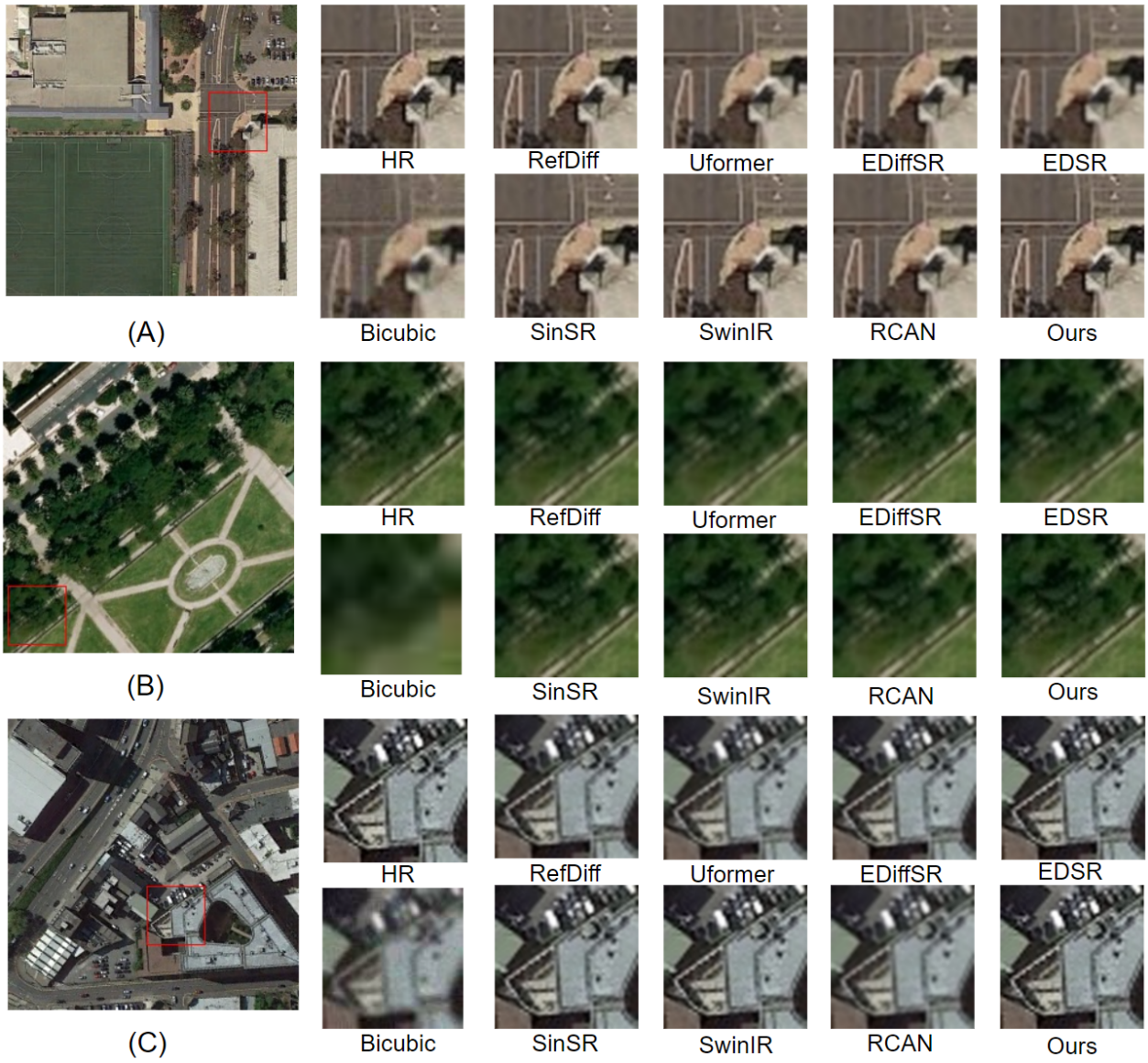

Figure 7. Qualitative comparison at ×3 scale under blind degradation on the AID dataset. Three groups of images are shown: (**A**) first group, (**B**) second group, (**C**) third group. For each group, results from different methods are presented, and the proposed approach demonstrates superior robustness and detail preservation against complex degradations involving blur and noise. The red square highlights the region of interest in each image for comparison.

4. Discussion

4.1. Ablation Study

To validate the contributions of our core modules, we conducted an ablation study on the AID remote sensing dataset. Our hierarchical multi-task restoration network consists of three modules: DAM, LGDF, and SRG. We started from a baseline network (Base) and progressively added one module at a time to quantify the performance improvements. Model configurations are as follows: Base: baseline UNet encoder-decoder; Base + DAM: baseline with DAM to extract degradation features; Base + DAM + LGDF: further addition of LGDF for complementary convolutional and Transformer decoding; Base + DAM + LGDF + SRG (Full Model): complete network including SRG.

As shown in Table 3 and Figure 8, the model improves progressively as modules are added. Incorporating DAM increases PSNR by 0.63 dB and SSIM by 0.011, demonstrating that explicit multi-scale degradation modeling effectively captures low-resolution image

priors and guides structure recovery in blurry or noisy scenes. Alternative designs such as implicit degradation embeddings or stochastic conditioning were tested in preliminary experiments but showed lower controllability and unstable structure reconstruction, motivating our choice of explicit modeling. Adding LGDF further increases PSNR to 28.05 dB and reduces LPIPS by 0.007, indicating that the combination of convolutional and Transformer decoders preserves global structural consistency while enhancing local textures. Designs using only convolutional decoders or only Transformer decoders resulted in either blurred textures or inconsistent global structures, supporting the necessity of the hybrid global-local decoding strategy. Finally, including SRG achieves the best performance (PSNR 28.45 dB, SSIM 0.837, LPIPS 0.115), as its structural-prior-based regularization stabilizes generation and suppresses artifacts. A version without SRG produced unstable edges and slight artifacts, justifying the inclusion of structural regularization.

Table 3. Ablation study on the AID dataset.

Method	DAM	LGDF	SRG	PSNR/SSIM/LPIPS
Base				27.15/0.812/0.138
Base + DAM	✓			27.78/0.823/0.130
Base + DAM + LGDF	✓	✓		28.05/0.809/0.123
Base + DAM + LGDG + SRG	✓	✓	✓	28.45/0.837/0.115

Overall, these results confirm that each module is not only effective individually but also that our specific design choices—explicit degradation modeling, global-local decoding, and structural regularization—are carefully chosen based on empirical comparisons and collectively enable superior pixel-level accuracy and perceptual quality compared to conventional alternatives.

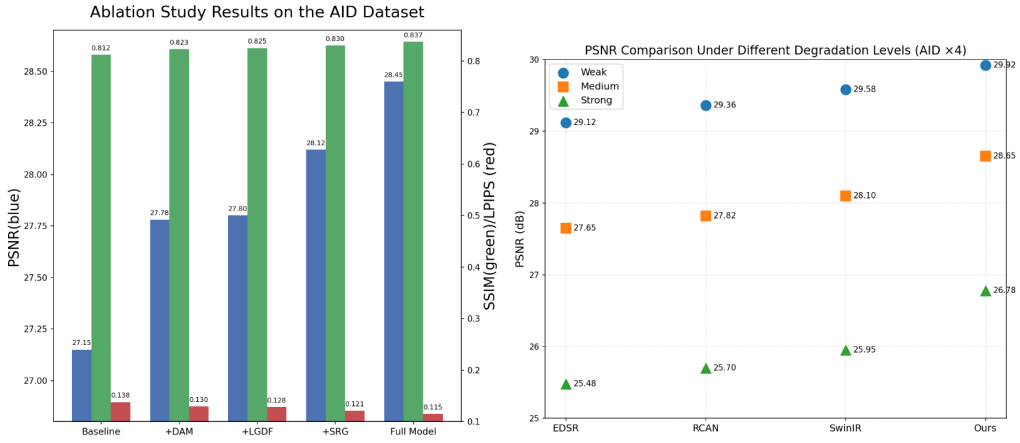

Figure 8. Ablation Study Results on the AID Dataset and PSNR Comparison under Different Degradation Levels on AID ×4.

4.2. Further Analysis: Robustness Under Different Degradation Levels

We evaluated the robustness of the proposed method against varying degradation conditions on the AID dataset, for the weak degradation (low noise and mild blur), medium degradation (moderate noise and blur), and strong degradation (high noise and severe blur). We compared the proposed approach with several representative super-resolution models such as EDSR, RCAN and SwinIR and measured PSNR variations under the ×4 upscaling task. Table 4 summarizes the experimental data. Overall, each method gets a small drop in performance. Our method gets smaller declines compared to competing approaches. For example, when a degradation goes from weak to strong, EDSR drops 3.64 dB and our method

drops 3.14 dB in PSNR. This indicates that our DAM can adaptively model degradation properties and achieve good restoration performance with complex degradations.

Table 4. PSNR comparison under different degradation levels (AID, ×4). The best results are highlighted in bold and colored in red to indicate important data.

Method	Weak	Medium	Strong	Drop
EDSR	29.12	27.65	25.48	−3.64
RCAN	29.36	27.82	25.70	−3.66
SwinIR	29.58	28.10	25.95	−3.63
Ours	**29.92**	**28.65**	**26.78**	**−3.14**

Moreover, the scatter plots in Figure 8 show the degradation trends more clearly. Competing methods drop sharply as degradation increases. In contrast, our method stays more stable, showing lower sensitivity and better robustness. Overall, the proposed method consistently outperforms others under different degradation levels. This further confirms the adaptive ability and generalization of the DAM module for handling complex degradations. Remote sensing imagery is inevitably affected by various imaging variabilities, including different sensor characteristics, noise levels, blur patterns, and atmospheric conditions. The above robustness analysis demonstrates that the proposed method can effectively handle such variabilities. By explicitly modeling degradation characteristics through the DAM module, the framework adapts its reconstruction strategy under different degradation levels instead of relying on fixed or idealized assumptions. As a result, the proposed method exhibits smaller performance degradation and more stable behavior compared with competing approaches, indicating strong robustness and generalization capability in realistic remote sensing scenarios.

4.3. Computational Efficiency Analysis

In addition to reconstruction accuracy, computational efficiency is an important factor for practical remote sensing applications. Table 5 presents the comparison of model parameters and FLOPs among different methods under the ×4 super-resolution setting. Although our method incorporates Transformer-based components, the overall model size remains comparable to recent diffusion-based SR methods. More importantly, by adopting a single-step diffusion formulation, the proposed framework avoids repeated forward passes during inference, resulting in significantly lower computational cost compared with conventional multi-step diffusion-based approaches. Therefore, the proposed method achieves a favorable balance between reconstruction quality and computational efficiency, making it suitable for large-scale and real-world RSISR tasks.

Table 5. Model complexity comparison under ×4 super-resolution.

Method	Params (M)	FLOPs (G)
EDSR	43.1	290.0
RCAN	15.6	330.0
SwinIR	11.9	215.0
Uformer	16.8	240.0
SinSR	19.4	260.0
RefDiff	22.7	285.0
EDiffSR	24.1	310.0
Ours	18.9	245.0

4.4. Limitations and Future Work

Despite the promising performance, the proposed method still has several limitations. First, although the single-step diffusion framework significantly reduces inference time

compared with conventional multi-step diffusion models, the incorporation of Transformer-based decoders introduces additional computational overhead compared with lightweight CNN-based methods. Further model compression or lightweight attention designs could be explored to improve efficiency. Second, while the degradation-aware modeling module improves robustness under mixed degradations, the degradation types considered in this work are still limited to commonly used blur and noise models. More complex real-world degradations, such as severe atmospheric distortions or sensor-specific artifacts, remain challenging and will be investigated in future work. Finally, the proposed framework focuses on single-image super-resolution. Extending the model to multi-temporal or multi-modal remote sensing data (e.g., incorporating SAR or multispectral information) is a promising direction to further enhance reconstruction quality and practical applicability.

5. Conclusions

In this work, we presented a degradation-aware diffusion-based framework for remote sensing image super-resolution, which integrates explicit degradation modeling, a dual-decoder reconstruction architecture, and static regularization-guided generation. By explicitly encoding diverse and complex degradation characteristics, the proposed degradation-aware module provides informative priors that effectively guide the diffusion process under challenging real-world conditions. Meanwhile, the dual-decoder structure collaboratively exploits convolutional and Transformer-based representations, enabling accurate structural reconstruction while preserving fine-grained texture details. Furthermore, the introduced static regularization guidance stabilizes the generation process and enhances structural consistency across different degradation levels and scaling factors. Extensive experiments conducted on multiple widely used remote sensing benchmark datasets demonstrate that the proposed method consistently outperforms state-of-the-art approaches in both objective metrics and visual quality. Ablation studies further verify the effectiveness of each individual component and reveal their complementary roles in improving robustness against diverse degradation conditions. Despite these promising results, there remain several directions for future research. First, the proposed framework can be extended to more challenging real-world remote sensing scenarios, such as multispectral and hyperspectral image super-resolution. Second, incorporating cross-modal priors, including geographic information and semantic annotations, may further enhance reconstruction realism and structural consistency. Overall, we believe that this work provides a solid methodological foundation for achieving high-quality and robust remote sensing image super-resolution.

Author Contributions: Conceptualization, X.C.; Methodology, C.D.; Formal analysis, T.F.; Investigation, J.H.; Data curation, J.H.; Writing—original draft, X.C. and C.D.; Writing—review and editing, X.C. and C.D.; Visualization, T.F.; Funding acquisition, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. 62301427, the Natural Science Basic Research Program of Shaanxi Province under Grant No. 2025JC-YBQN-906, the Key Projects for Research and Development of Shaanxi Province under Grant No. 2025CY-YBXM-067, the Education Department of Shaanxi Province under Grant No. 24JK0651 and the Key Scientific Research Projects of the Shaanxi Provincial Education Department under Grant No. 24JR152, the National Natural Science Foundation of China (No. 62577044).

Data Availability Statement: The datasets used in this study are publicly available. The AID dataset is available from its official website (https://captain-whu.github.io/AID/, accessed on 14 October 2025), and the UCMerced Land Use Dataset can be obtained from the official site (https://vision.ucmerced.edu/datasets/, accessed on 14 October 2025).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40.
2. Li, Y.; Qi, F.; Wan, Y. Improvements on bicubic image interpolation. In Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20–22 December 2019; Volume 1, pp. 1316–1320.
3. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* **2003**, *29*, 1153–1160. [CrossRef]
4. Dong, W.; Zhang, L.; Shi, G.; Wu, X. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Trans. Image Process.* **2011**, *20*, 1838–1857. [CrossRef] [PubMed]
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
6. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
7. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
8. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17683–17693.
9. Pereira, G.A.; Hussain, M. A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. *arXiv* **2024**, arXiv:2408.15178. [CrossRef]
10. Wang, X.; Yi, J.; Guo, J.; Song, Y.; Lyu, J.; Xu, J.; Yan, W.; Zhao, J.; Cai, Q.; Min, H. A review of image super-resolution approaches based on deep learning and applications in remote sensing. *Remote Sens.* **2022**, *14*, 5423.
11. Yang, D.; Li, Z.; Xia, Y.; Chen, Z. Remote sensing image super-resolution: Challenges and approaches. In Proceedings of the 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 21–24 July 2015; pp. 196–200.
12. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
13. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
14. Zhang, N.; Wang, Y.; Zhang, X.; Xu, D.; Wang, X.; Ben, G.; Zhao, Z.; Li, Z. A multi-degradation aided method for unsupervised remote sensing image super resolution with convolution neural networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 1–14.
15. Rudin, L.I.; Osher, S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. D Nonlinear Phenom.* **1992**, *60*, 259–268. [CrossRef]
16. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4713–4726. [CrossRef]
17. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
18. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502.
19. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 694–711.
20. Ho, J.; Saharia, C.; Chan, W.; Fleet, D.J.; Norouzi, M.; Salimans, T. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.* **2022**, *23*, 1–33.
21. Yue, Z.; Wang, J.; Loy, C.C. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 13294–13307.
22. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8162–8171.
23. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1905–1914.
24. Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.P.; Liu, Z.; Qiao, Y.; Kot, A.C.; Wen, B. Sinsr: Diffusion-based image super-resolution in a single step. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 25796–25805.

25. Dong, R.; Yuan, S.; Luo, B.; Chen, M.; Zhang, J.; Zhang, L.; Li, W.; Zheng, J.; Fu, H. Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27684–27694.
26. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Jin, X.; Zhang, L. EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 1–14.
27. Zhang, K.; Liang, J.; Van Gool, L.; Timofte, R. Designing a practical degradation model for deep blind image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4791–4800.
28. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1664–1673.
29. Yue, Z.; Zhao, Q.; Xie, J.; Zhang, L.; Meng, D.; Wong, K.Y.K. Blind image super-resolution with elaborate degradation modeling on noise and kernel. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2128–2138.
30. Wang, L.; Wang, Y.; Dong, X.; Xu, Q.; Yang, J.; An, W.; Guo, Y. Unsupervised degradation representation learning for blind super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 10581–10590.
31. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
32. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
33. Dong, R.; Mou, L.; Zhang, L.; Fu, H.; Zhu, X.X. Real-world remote sensing image super-resolution via a practical degradation model and a kernel-aware network. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 155–170.
34. Zhang, J.; Xu, T.; Li, J.; Jiang, S.; Zhang, Y. Single-image super resolution of remote sensing images with real-world degradation modeling. *Remote Sens.* **2022**, *14*, 2895.
35. Qin, Y.; Nie, H.; Wang, J.; Liu, H.; Sun, J.; Zhu, M.; Lu, J.; Pan, Q. Multi-degradation super-resolution reconstruction for remote sensing images with reconstruction features-guided kernel correction. *Remote Sens.* **2024**, *16*, 2915.
36. Liang, J.; Zeng, H.; Zhang, L. Efficient and degradation-adaptive network for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 574–591.
37. Aybar, C.; Montero, D.; Contreras, J.; Donike, S.; Kalaitzis, F.; Gómez-Chova, L. SEN2NAIP: A large-scale dataset for Sentinel-2 Image Super-Resolution. *Sci. Data* **2024**, *11*, 1389.
38. Zhu, H.; Tang, X.; Xie, J.; Song, W.; Mo, F.; Gao, X. Spatio-temporal super-resolution reconstruction of remote-sensing images based on adaptive multi-scale detail enhancement. *Sensors* **2018**, *18*, 498. [CrossRef]
39. Wang, Y.; Shao, Z.; Lu, T.; Huang, X.; Wang, J.; Zhang, Z.; Zuo, X. Lightweight remote sensing super-resolution with multi-scale graph attention network. *Pattern Recognit.* **2025**, *160*, 111178. [CrossRef]
40. Chen, Y.; Zhang, X. Ddsr: Degradation-aware diffusion model for spectral reconstruction from rgb images. *Remote Sens.* **2024**, *16*, 2692. [CrossRef]
41. Wang, Z.; Xia, M.; Weng, L.; Hu, K.; Lin, H. Dual encoder–decoder network for land cover segmentation of remote sensing image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 2372–2385.
42. Kim, S.P.; Su, W.Y. Recursive high-resolution reconstruction of blurred multiframe images. *IEEE Trans. Image Process.* **1993**, *2*, 534–539. [CrossRef] [PubMed]
43. Zhang, X.; Zhu, K.; Chen, G.; Tan, X.; Zhang, L.; Dai, F.; Liao, P.; Gong, Y. Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network. *Remote Sens.* **2019**, *11*, 755. [CrossRef]
44. Gao, H.; Zhang, Y.; Yang, J.; Dang, D. Mixed hierarchy network for image restoration. *Pattern Recognit.* **2025**, *161*, 111313. [CrossRef]
45. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.S.; Khan, F.S. Transformers in remote sensing: A survey. *Remote Sens.* **2023**, *15*, 1860.
46. Zafar, A.; Aftab, D.; Qureshi, R.; Fan, X.; Chen, P.; Wu, J.; Ali, H.; Nawaz, S.; Khan, S.; Shah, M. Single stage adaptive multi-attention network for image restoration. *IEEE Trans. Image Process.* **2024**, *33*, 2924–2935. [CrossRef]
47. Chung, H.; Kim, J.; Mccann, M.T.; Klasky, M.L.; Ye, J.C. Diffusion posterior sampling for general noisy inverse problems. *arXiv* **2022**, arXiv:2209.14687.
48. Ng, M.K.; Shen, H.; Lam, E.Y.; Zhang, L. A total variation regularization based super-resolution reconstruction algorithm for digital video. *EURASIP J. Adv. Signal Process.* **2007**, *2007*, 074585. [CrossRef]

49. Ma, C.; Rao, Y.; Cheng, Y.; Chen, C.; Lu, J.; Zhou, J. Structure-preserving super resolution with gradient guidance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7769–7778.
50. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
51. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981.
52. Jähne, B. *Digital Image Processing*; Springer: Berlin/Heidelberg, Germany, 2005.
53. Sheikh, H.R.; Bovik, A.C. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [CrossRef]