

Comparative Study of Stable Diffusion-Based Super-Resolution Methods for Remote Sensing Imagery

Yiran Lu^{1,2}, Hongjie He^{1,2,3}, Fei Zhu^{1,2}, Jonathan Li^{*1,2,3}

¹School of GeoAI and Hinton STAI Institute, East China Normal University,
Shanghai, 200241, China - 52290161012@stu.ecnu.edu.cn

²Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University,
Shanghai, 200241, China - jli@geoai.ecnu.edu.cn

³Department of Geography and Environmental Management, University of Waterloo,
Waterloo, ON N2L 3G1, Canada

Keywords: Stable Diffusion, Super-resolution, Remote Sensing, Comparative Study, Diffusion-based Super-Resolution.

Abstract

Remote sensing image super-resolution aims to recover fine structural and textural details from degraded low-resolution observations. However, conventional methods and early deep learning models often produce over-smoothed results and struggle to reconstruct realistic high-frequency content. Stable Diffusion-based (SD-based) methods offer a promising alternative by using strong generative priors to synthesize more natural, detail-rich super-resolved images. Although many SD-based super-resolution methods have been proposed in computer vision, their use in remote sensing imagery remains limited, and systematic comparative evaluation in this domain is still lacking, leaving insufficient empirical guidance for method development. Therefore, this paper compares four representative SD-based super-resolution methods, namely Stable Super-Resolution (StableSR), Semantics-Aware Super-Resolution (SeeSR), Different Blind Image Restoration (DiffBIR), and Pixel-Aware Stable Diffusion (PASD), on the WHU-Mix remote sensing dataset. The evaluation uses seven metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Frechet Inception Distance (FID), CLIP Image Quality Assessment (CLIP-IQA), Multi-Scale Image Quality Transformer (MUSIQ), and Multi-Dimension Attention Network for No-Reference Image Quality Assessment (MANIQA). Quantitative results show that StableSR achieves the highest PSNR of 23.16 dB, PASD obtains the best SSIM of 0.81 and lowest LPIPS of 0.45, SeeSR achieves the best MUSIQ of 64.57 and MANIQA of 0.46, and DiffBIR achieves the best FID of 110.58 and CLIP-IQA of 0.68 but with weaker full-reference fidelity. These findings indicate that current SD-based methods favor different aspects, including fidelity preservation, perceptual quality, and generative realism, and should be selected according to the target remote sensing application.

1. Introduction

High-spatial-resolution remote sensing imagery is of great importance for a wide range of applications, including urban planning, land-use monitoring, environmental assessment, disaster response, and precision agriculture. In these scenarios, fine-grained spatial details are essential for accurately identifying ground objects, delineating structural boundaries, and supporting subsequent interpretation or decision-making processes. However, acquiring high-resolution remote sensing images directly is often constrained by sensor hardware, imaging altitude, atmospheric conditions, revisit cycles, and economic cost. As a result, remote sensing image super-resolution (SR), which aims to reconstruct high-resolution (HR) images from low-resolution (LR) observations, has become an important and practical research topic.

Despite its significance, remote sensing super-resolution remains highly challenging. Compared with natural images, remote sensing images usually contain more complex spatial distributions, larger scale variations, dense and repetitive textures, and diverse land-cover categories within the same scene. Fine structures such as roads, building edges, field boundaries, and small targets are often easily degraded during image acquisition, and different local regions within the same image may exhibit different degradation characteristics, such as blur, noise, compression artifacts, or aliasing introduced by downsampling. These characteristics make it difficult for SR models to restore

visually realistic details while preserving geometric structure and semantic consistency. In addition, the objective of remote sensing SR is not limited to improving visual sharpness alone. The reconstructed images are often expected to remain reliable for downstream tasks such as object detection, scene interpretation, and change analysis. Therefore, a satisfactory SR method should not only enhance perceptual quality, but also avoid over-smoothing important textures or generating misleading details. This requirement highlights the need for methods that can better balance reconstruction fidelity and perceptual quality in remote sensing imagery.

To address the remote sensing SR problem, a variety of methods have been proposed, ranging from interpolation-based and reconstruction-based approaches to deep learning models. Early traditional methods are simple and computationally efficient, but they usually rely on handcrafted priors and limited assumptions about image degradation, making them insufficient for recovering complex high-frequency details in remote sensing scenes. With the development of deep learning, convolutional neural network-based (CNN-based) and generative adversarial network-based (GAN-based) super-resolution methods have significantly improved reconstruction capability and visual quality and have gradually become mainstream solutions for image SR. Nevertheless, these methods still exhibit evident limitations when applied to remote sensing imagery. Most distortion-oriented SR models are optimized mainly for pixel-

wise fidelity, such as minimizing mean squared error or directly optimizing reconstruction accuracy, and therefore tend to produce overly smooth results with weak texture representation. Although such methods may achieve favorable quantitative scores, they often fail to reconstruct realistic fine-grained details of ground objects, especially in regions containing dense vegetation, building boundaries, road networks, and other complex structures. GAN-based methods alleviate over-smoothing by introducing perceptual and adversarial learning and can generate sharper textures, but in remote sensing scenarios they often produce unstable local details or hallucinated patterns that do not faithfully correspond to true surface content. Such errors are particularly problematic because they can directly affect scene interpretation and subsequent application-oriented analysis. In summary, previous super-resolution methods have improved image quality to a certain extent, but they still face substantial challenges in balancing reconstruction fidelity and perceptual quality in remote sensing imagery. These limitations motivate the exploration of more powerful generative priors and more robust restoration frameworks for remote sensing SR.

Recently, diffusion-based generative models, especially Stable Diffusion (SD) (Rombach et al., 2022), have provided a new framework for image super-resolution and restoration. Benefiting from large-scale generative priors learned from diverse image data, Stable Diffusion-based (SD-based) methods can synthesize more natural, detailed, and semantically meaningful visual content than many conventional SR approaches. Compared with distortion-oriented models, they exhibit stronger capability in restoring realistic textures; compared with GAN-based methods, they often provide more stable generation behavior and a more flexible trade-off between perceptual quality and fidelity. These properties make SD-based methods particularly promising for remote sensing SR, where the recovery of fine structures and land-cover details is crucial.

Building upon the Stable Diffusion framework, several representative super-resolution methods have recently been proposed, including Stable Super-Resolution (StableSR), Semantics-Aware Super-Resolution (SeeSR), Different Blind Image Restoration (DiffBIR), and Pixel-Aware Stable Diffusion (PASD). Although all of them exploit the generative prior of diffusion models, they differ substantially in design philosophy and restoration emphasis. StableSR emphasizes time-aware control of image guidance and controllable fidelity-realism balancing. SeeSR introduces degradation-aware semantic prompting to improve semantic consistency during restoration. DiffBIR adopts a two-stage blind restoration framework that decouples degradation removal and content regeneration. PASD combines pixel-aware structural guidance with semantic prompting to preserve local spatial correspondence. These differences suggest that SD-based methods may exhibit distinct advantages and limitations when applied to remote sensing imagery.

However, despite their encouraging results in natural-image super-resolution and general image restoration, the performance of these SD-based methods in the remote sensing domain has not yet been investigated in a sufficiently systematic manner. Existing studies remain limited, and a comprehensive comparison under a unified experimental setting is still lacking, especially with regard to how different SD-based methods behave on remote sensing scenes with diverse textures, structures, and degradation characteristics. As a result, it remains unclear which method is more suitable for semantic preservation, which is more robust to complex degradation, and which

can better balance perceptual quality with structural fidelity in remote sensing SR tasks.

To address this gap, the present study establishes a unified comparative evaluation on the WHU-Mix remote sensing dataset. A common evaluation framework is used to assess the four representative SD-based methods from multiple perspectives, including full-reference fidelity, perceptual similarity, distribution-level realism, and no-reference image quality. Through quantitative and qualitative analysis, the study aims to reveal the strengths and limitations of different SD-based methods in remote sensing super-resolution and to provide useful insights for future research and practical applications in this field.

The main contributions of this study are threefold. First, a unified comparative framework is established for evaluating representative SD-based super-resolution methods on remote sensing imagery under a realistic $\times 4$ degradation setting. Second, four representative SD-based methods, namely StableSR, SeeSR, DiffBIR, and PASD, are systematically analyzed from the perspectives of model design, conditional guidance strategy, and restoration objective. Third, comprehensive experiments reveal clear differences between fidelity-oriented and perceptual-oriented restoration behavior, thereby providing application-oriented guidance for selecting SD-based super-resolution methods in remote sensing scenarios.

2. Stable Diffusion-Based Super-Resolution Methods

SD-based super-resolution methods formulate restoration as a conditional denoising process in latent space. Compared with traditional super-resolution models optimized mainly for pixel-wise distortion, these methods exploit pretrained generative priors to recover richer textures and more realistic high-frequency details. Their core advantage lies in the ability to synthesize visually plausible structures under complex degradation, but this advantage is accompanied by a core risk: generated details must remain faithful to the degraded observation rather than drifting toward semantically or geometrically incorrect content.

This issue is particularly important in remote sensing imagery. Unlike many natural-image restoration scenarios, remote sensing super-resolution is closely related to practical interpretation tasks such as building extraction, road delineation, land-cover recognition, and change analysis. Consequently, a high-performing remote sensing SR method should not only improve perceptual sharpness, but also preserve object boundaries, structural layouts, and scene semantics. These requirements make remote sensing restoration an especially suitable testbed for comparing different SD-based design strategies.

Existing SD-based SR methods can be distinguished by three major dimensions: the form of conditional guidance, the target of restoration control, and the way fidelity-perceptual-quality trade-offs are handled. Some methods strengthen low-level structural guidance to preserve geometry, while others inject semantic information to reduce ambiguous generation. Some approaches are designed for blind and mixed degradations, whereas others aim to improve controllability within the denoising process itself.

2.1 StableSR

StableSR (Wang et al., 2024) is designed to exploit the generative prior of Stable Diffusion for real-world super-resolution

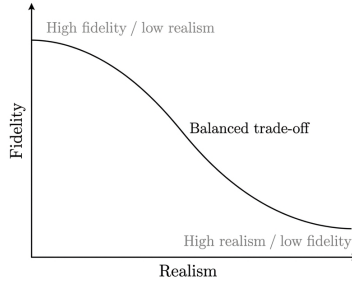


Figure 1. Conceptual illustration of the fidelity-realism trade-off in diffusion-based super-resolution. Increasing realism often improves visual richness but may reduce strict reconstruction fidelity, whereas stronger fidelity constraints may suppress generative flexibility.

while maintaining a controllable balance between reconstruction fidelity and visual realism. Its core idea is that the strength of image guidance should vary across denoising steps rather than remain fixed. The time-aware encoder therefore modulates conditional information according to the diffusion timestep, allowing the model to provide stronger structural guidance when necessary while preserving the flexibility of diffusion-based generation.

Another important component is the controllable feature wrapping (CFW) module, which explicitly regulates the fusion between encoder and decoder features in the autoencoder branch. Through this mechanism, StableSR can adjust the trade-off between fidelity and realism in a more controllable manner. In addition, progressive aggregation sampling enables patch-wise restoration for larger images and alleviates the fixed-resolution limitation of pretrained Stable Diffusion models.

The CFW operation can be written as

$$\mathbf{F}_m = \mathbf{F}_d \mathcal{C}(\mathbf{F}_e, \mathbf{F}_d; \theta) \times w, \quad (1)$$

where $\mathbf{F}_e$ and $\mathbf{F}_d$ denote the encoder and decoder features, respectively, $\mathbf{F}_m$ denotes the modulated decoder feature, $\mathcal{C}(\cdot; \theta)$ represents trainable convolution layers, and $w \in [0, 1]$ is an adjustable coefficient. This formulation directly reflects how StableSR controls the fidelity-realism trade-off through CFW.

The fidelity-realism trade-off controlled by CFW is conceptually illustrated in Figure 1. From a design perspective, StableSR can be regarded as a diffusion framework centered on time-aware control and controllable fidelity-realism balancing. The overall StableSR framework is illustrated in Figure 2.

2.2 SeeSR

SeeSR (Wu et al., 2024) focuses on semantic preservation in real-world super-resolution. Its design is motivated by the observation that low-resolution inputs often contain semantic ambiguity, especially under severe degradations. Under such conditions, diffusion models may generate visually plausible textures that nevertheless correspond to incorrect object categories or scene semantics. To reduce this risk, SeeSR introduces degradation-aware semantic prompting into the restoration framework.

Its core module, the degradation-aware prompt extractor (DAPE), produces both hard prompts and soft semantic em-

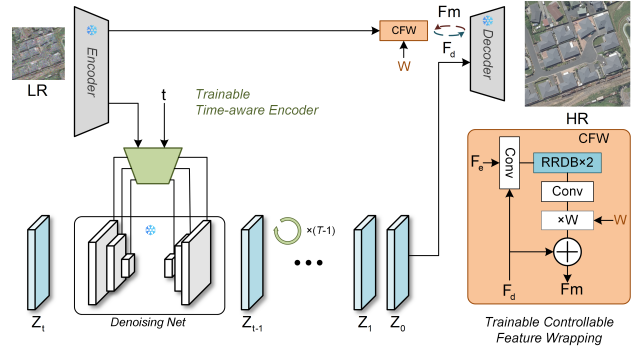


Figure 2. Overview of the StableSR framework. The method combines a time-aware encoder with controllable feature wrapping to regulate the balance between reconstruction fidelity and visual realism during diffusion-based restoration.

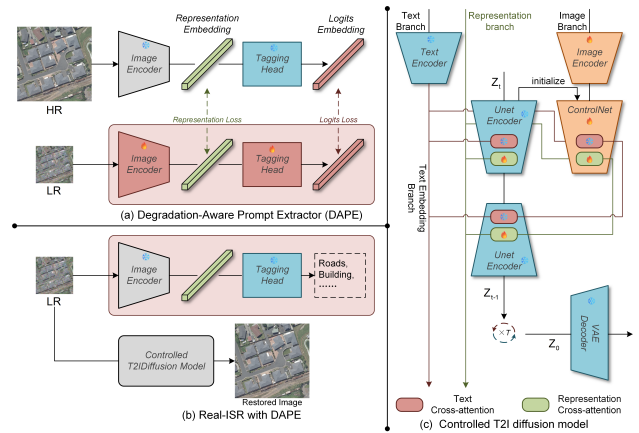


Figure 3. Overview of the SeeSR framework. The method introduces a degradation-aware prompt extractor and a controlled text-to-image diffusion model to improve semantic consistency in real-world super-resolution.

beddings from degraded inputs. These semantic cues are injected into a controlled text-to-image diffusion model together with the image condition, so that generation is constrained not only by local image evidence but also by higher-level scene understanding. During inference, low-resolution embedding refinement is further used to reduce inconsistencies between the degraded input and the starting latent state.

The training objective of DAPE is written as

$$\mathcal{L}_{\text{DAPE}} = \mathcal{L}_r(f_y^{\text{rep}}, f_x^{\text{rep}}) + \lambda \mathcal{L}_l(f_y^{\text{logits}}, f_x^{\text{logits}}), \quad (2)$$

where x and y denote the HR image and its degraded LR counterpart, respectively, f_y^{rep} and f_x^{logits} denote the representation embedding and logits embedding, and λ is a balancing coefficient. This loss aligns semantic representations extracted from degraded inputs with those of clean images, which is the key to SeeSR's semantics-aware restoration.

This semantics-oriented design directly targets semantic ambiguity in degraded inputs. At the same time, its effectiveness depends strongly on the reliability of prompt estimation, which may be challenged by severe degradation, repetitive textures, and small targets. The overall pipeline of SeeSR is shown in Figure 3.

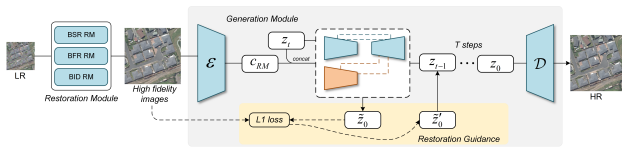


Figure 4. Overview of the DiffBIR framework. The method decouples blind image restoration into degradation removal and diffusion-based information regeneration, and further introduces region-adaptive restoration guidance to balance fidelity and realism during sampling.

2.3 DiffBIR

DiffBIR (Lin et al., 2024) is designed as a unified blind image restoration framework for unknown and mixed degradations. Instead of directly conditioning diffusion generation on severely degraded observations, it decouples the problem into two stages: degradation removal and information regeneration. In the first stage, a restoration module is used to suppress degradation artifacts and provide a cleaner condition image. In the second stage, missing content is regenerated using pretrained diffusion priors.

The generation stage is implemented through IRControlNet, which injects the intermediate restoration result into the diffusion backbone. In addition, DiffBIR introduces a training-free region-adaptive restoration guidance mechanism during inference, enabling low-frequency regions to follow the high-fidelity condition more closely while allowing high-frequency regions to retain greater generative flexibility. Such a design is intended to improve controllability under blind and mixed degradations.

Its region-adaptive restoration guidance can be expressed as

$$\mathcal{L}(\tilde{z}_0) = \frac{1}{HW C} \|\mathcal{W} \odot (\mathcal{D}(\tilde{z}_0) - I_{RM})\|_2^2, \quad (3)$$

where $\tilde{z}_0$ is the predicted clean latent, $\mathcal{D}(\cdot)$ is the latent decoder, I_{RM} is the stage-I restoration result, and $\mathcal{W}$ is a spatially adaptive weighting map. This formulation shows that DiffBIR does not impose uniform guidance across the image, but instead regulates fidelity with region-dependent strength during sampling.

The overall DiffBIR framework is illustrated in Figure 4. Compared with more directly conditioned methods, DiffBIR emphasizes decoupled restoration and robustness rather than strict pixel-level preservation. Its final performance, however, remains dependent on the quality of the first-stage restoration and is subject to error propagation across stages.

2.4 PASD

PASD (Yang et al., 2024) places particular emphasis on pixel-level structural preservation in real-world super-resolution. Its framework includes a degradation removal module, a high-level information extraction module, an adjustable noise schedule, and a pixel-aware cross-attention module. Its key idea is that diffusion priors should be constrained not only by global image conditions but also by explicit local correspondence between degraded inputs and restored outputs. To achieve this goal, PASD introduces pixel-aware cross-attention (PACA), which injects local spatial cues into the denoising process and improves the preservation of edges, contours, and texture layouts.

PASD further combines degradation-insensitive structural guidance with an adjustable noise schedule (ANS), so that the

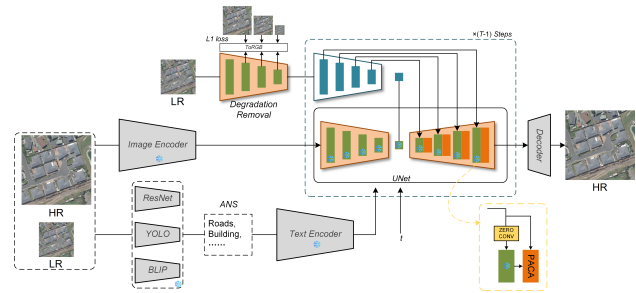


Figure 5. Overview of the PASD framework. The method combines degradation removal, high-level semantic guidance, adjustable noise scheduling, and pixel-aware cross-attention to preserve local structure while improving perceptual realism.

sampling process can better balance geometric faithfulness and perceptual enhancement. Although high-level semantic information is also incorporated, the method remains primarily structure-oriented rather than semantics-oriented.

The core PACA operation is formulated as

$$\text{PACA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \right) \mathbf{V}, \quad (4)$$

where $\mathbf{Q}$ is computed from latent diffusion features, while $\mathbf{K}$ and $\mathbf{V}$ are derived from condition features carrying pixel-level spatial structure. This attention design enables PASD to preserve local geometric correspondence more explicitly than standard global condition injection.

The overall PASD framework is illustrated in Figure 5. From a comparative perspective, PASD can be regarded as a diffusion SR method centered on pixel-aware structural guidance and controllable perceptual-fidelity balancing.

2.5 Comparative Analysis of SD-Based Super-Resolution Methods

Although the four methods share the same Stable Diffusion foundation, they correspond to markedly different restoration priorities. StableSR emphasizes time-aware control of image guidance and fidelity-realism balancing. SeeSR emphasizes semantic preservation through degradation-aware prompt learning. DiffBIR improves robustness by decoupling degradation removal and content regeneration in a two-stage framework. PASD directly targets local geometric consistency through pixel-aware structural guidance. These differences are not merely implementation details, but reflect distinct assumptions about what should be preserved during restoration.

For remote sensing super-resolution, the distinction among these design priorities is especially meaningful. Geometric precision is critical for man-made structures and boundaries, semantic consistency is important for scene interpretation, and robustness to mixed degradations is necessary for realistic observation conditions. As a result, no single method can be assumed to dominate in all settings. A comparative study should therefore examine not only aggregate metric ranking, but also the relationship between model design and quantitative performance characteristics.

Based on the above considerations, the subsequent experiments focus on three analytical dimensions: reconstruction fidelity, perceptual quality, and robustness under realistic degradation.

Through this framework, the relative strengths and limitations of StableSR, SeeSR, DiffBIR, and PASD can be interpreted through clearer analytical criteria rather than as isolated numerical results.

To make the methodological differences clearer, Table 1 summarizes the guidance strategy, core design, design objective, and intrinsic limitation of the four compared methods. This summary table is intended to clarify their methodological differences at the design level.

3. Experiment and Results Analyses

3.1 Dataset Description

To evaluate the performance of different SD-based super-resolution methods on remote sensing images, this study conducts experiments on the WHU-Mix dataset (Luo et al., 2022). WHU-Mix is a publicly available large-scale remote sensing building dataset originally developed for building extraction and related generalization studies. According to the original dataset description, it consists of two parts: a training/validation (trainval) subset and a test subset. The trainval subset contains 43,727 image patches of 512×512 pixels and was constructed from two existing datasets together with newly edited and newly acquired samples, in order to increase scene diversity. The test subset contains 8,402 image patches of the same size, as well as original uncropped remote sensing images collected from five cities on five continents. Representative examples are shown in Figure 6.

This organization is important for the present study because the dataset provides both globally distributed patch-based samples for model development and geographically separated test data for final evaluation. There is no geographic overlap between the trainval subset and the test subset. In addition, the spatial resolution spans a wide range from 0.09 to 2.5 m/pixel, reflecting substantial cross-region and cross-sensor variation. The imagery covers diverse scene elements, including buildings, roads, vegetation, bare land, and mixed land-cover patterns, which makes the dataset suitable for evaluating restoration methods under realistic remote sensing conditions.



Figure 6. Examples from the WHU-Mix dataset. The first, second, and third columns show representative residential, grassland, and road scenes, respectively.

For the super-resolution experiments in this paper, the original high-resolution WHU-Mix images are used as reference images, and the corresponding low-resolution inputs are synthesized according to the degradation setting described in the next

subsection. Following the official split of WHU-Mix, 39,346 image patches are used for training, 4,381 image patches are used for validation, and 8,402 image patches are used for testing. Thus, the 43,727-image trainval subset is divided into training and validation portions, while the 8,402-image test subset is reserved for final quantitative and qualitative evaluation. Since all compared methods adopt 512×512 image patches in the final experimental protocol, the dataset can be used directly without additional patch resizing at the dataset level.

Overall, the WHU-Mix dataset provides a representative and challenging experimental basis for remote sensing super-resolution. Its global coverage, wide resolution span, and diverse scene composition enable a more comprehensive evaluation of the strengths and limitations of StableSR, SeeSR, DiffBIR, and PASD in remote sensing restoration tasks.

3.2 Degradation Setting and Task Formulation

The experiments are conducted under a unified $\times 4$ remote sensing super-resolution setting. Rather than adopting idealized bicubic downsampling only, the compared methods are evaluated under realistic synthetic degradations derived from the Real-ESRGAN paradigm (Wang et al., 2021), which itself builds upon practical blind super-resolution degradation modeling (Zhang et al., 2021). The degradation family includes mixed blur, random resizing, Gaussian or Poisson noise, JPEG compression, and sinc filtering, which better reflects the complex corruption patterns often observed in practical remote sensing imagery.

Although the four methods share the same task scale, their conditional inputs are not constructed in exactly the same way. PASD, StableSR, and DiffBIR rely on online or training-time degradation, while SeeSR uses offline prepared bicubic-upsampled conditional images generated from Real-ESRGAN-style degraded samples. The comparison therefore preserves methodological characteristics of the original implementations while maintaining a consistent $\times 4$ restoration objective across all methods.

3.3 Implementation Details for Reproducibility

For reproducibility, all compared methods are trained from Stable Diffusion 2.1 based checkpoints with a learning rate of 5×10^{-5} and an overall training schedule of 200,000 optimization steps. All compared methods use 512×512 training patches. Mixed precision, gradient accumulation, and checkpoint intervals follow the official settings of each method.

The main implementation differences lie in three aspects. First, the compared methods adopt different degradation pipelines: StableSR, PASD, and DiffBIR use online or training-time degradation generation, whereas SeeSR uses offline prepared bicubic-upsampled condition images derived from Real-ESRGAN-style degraded samples. Second, the construction of conditional inputs is not identical across methods, particularly in whether degraded images are resized back to the target resolution before diffusion conditioning. Third, the conditional signal is injected into the diffusion backbone in different forms, including time-aware image guidance, semantic prompting, two-stage restoration guidance, and pixel-aware structural guidance. These implementation differences are retained because they are intrinsic to the original method designs and directly influence restoration behavior.

Table 1. Design-level comparison of the compared SD-based super-resolution methods.

Method	Guidance strategy	Core design	Design objective	Intrinsic limitation
StableSR	Time-aware image guidance.	Time-aware encoder; controllable feature wrapping; progressive aggregation sampling.	Exploit diffusion prior while balancing reconstruction fidelity and visual realism.	Limited high-level semantic control from low-resolution-only conditioning.
SeeSR	Degradation-aware semantic prompting.	DAPE; controlled text-to-image diffusion; low-resolution embedding refinement.	Inject degradation-aware semantic priors to improve semantic consistency in real-world super-resolution.	Sensitive to prompt estimation quality under severe degradation or small-object scenes.
DiffBIR	Two-stage restoration with restoration guidance.	Restoration module for degradation removal; IRControlNet for diffusion regeneration; region-adaptive guidance.	Decouple degradation removal and content regeneration under blind degradations.	Dependent on stage-I restoration quality; susceptible to error propagation across stages.
PASD	Pixel-aware structural guidance with semantic prompts.	Degradation removal module; PACA; ANS; high-level information injection.	Preserve pixel-level structure while improving perceptual realism.	Coupled multi-module design; more constrained texture generation than realism-oriented methods.

3.4 Evaluation Metrics

To comprehensively assess the restoration performance of different SD-based super-resolution methods, this study adopts a multi-dimensional evaluation protocol including full-reference, perceptual, distribution-level, and no-reference quality metrics. Full-reference metrics include Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) (Hore and Ziou, 2010, Wang et al., 2004), which measure the numerical fidelity and structural consistency between the restored image I and the reference HR image I^{ref} . Their standard forms are written as

$$\text{PSNR}(I, I^{\text{ref}}) = 10 \log_{10} \left(\frac{L^2}{\frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W (I_{ij} - I_{ij}^{\text{ref}})^2} \right), \quad (5)$$

$$\text{SSIM}(I, I^{\text{ref}}) = \frac{(2\mu_I \mu_{I^{\text{ref}}} + c_1)(2\sigma_{II^{\text{ref}}} + c_2)}{(\mu_I^2 + \mu_{I^{\text{ref}}}^2 + c_1)(\sigma_I^2 + \sigma_{I^{\text{ref}}}^2 + c_2)}, \quad (6)$$

where L denotes the dynamic range, and μ , σ^2 , and $\sigma_{II^{\text{ref}}}$ denote the mean, variance, and covariance terms, respectively. PSNR is measured in dB. In general, higher PSNR and SSIM values indicate better reconstruction fidelity.

Perceptual similarity is evaluated by Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), which measures the feature-level distance between restored and reference images:

$$\text{LPIPS}(I, I^{\text{ref}}) = \left\| \Phi(I) - \Phi(I^{\text{ref}}) \right\|_2^2, \quad (7)$$

where $\Phi(\cdot)$ denotes the deep perceptual feature representation extracted by a pretrained network. A lower LPIPS value indicates that the restored image is perceptually closer to the ground truth. In addition, Fréchet Inception Distance (FID) (Saxena and Cao, 2021) is employed to assess the distribution-level realism of the generated results:

$$\text{FID}(X, Y) = \left\| \mu_X - \mu_Y \right\|_2^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{1/2} \right), \quad (8)$$

where (μ_X, Σ_X) and (μ_Y, Σ_Y) are the mean and covariance of Inception features extracted from generated and reference

image sets, respectively. Lower FID values suggest that the restored image distribution is more similar to that of real HR remote sensing images.

To further evaluate perceptual quality without relying on references, three no-reference image quality assessment metrics are adopted, namely CLIP Image Quality Assessment (CLIP-IQA) (Wang et al., 2023), Multi-Scale Image Quality Transformer (MUSIQ) (Ke et al., 2021), and Multi-Dimension Attention Network for Image Quality Assessment (MANIQA) (Yang et al., 2022). These metrics are computed by their corresponding pretrained evaluators and directly output quality scores for restored images. They provide complementary perspectives on visual naturalness, perceptual preference, and overall image quality. For these three metrics, higher values indicate better image quality. By combining the above metrics, the evaluation in this paper covers both reconstruction fidelity and perceptual quality, enabling a more comprehensive comparison of the strengths and limitations of the studied methods.

3.5 Quantitative Results

Table 2 reports the full-reference results, and Table 3 summarizes perceptual and no-reference metrics. The four methods do not follow a single ranking across all criteria, indicating different optimization priorities in fidelity, perceptual quality, and realism.

Table 2. Full-reference quantitative comparison on the WHU-Mix test set. Best results are in bold and second-best results are underlined.

Method	PSNR (dB)↑	SSIM↑	LPIPS↓
StableSR	23.16	<u>0.80</u>	0.49
SeeSR	21.98	0.75	<u>0.47</u>
PASD	<u>23.11</u>	0.81	0.45
DiffBIR	19.20	0.68	0.54

StableSR achieves the highest PSNR of 23.16 dB, whereas PASD obtains the best SSIM of 0.81 and the lowest LPIPS of 0.45. The two methods are therefore closest in full-reference fidelity, but their behaviors are different: StableSR

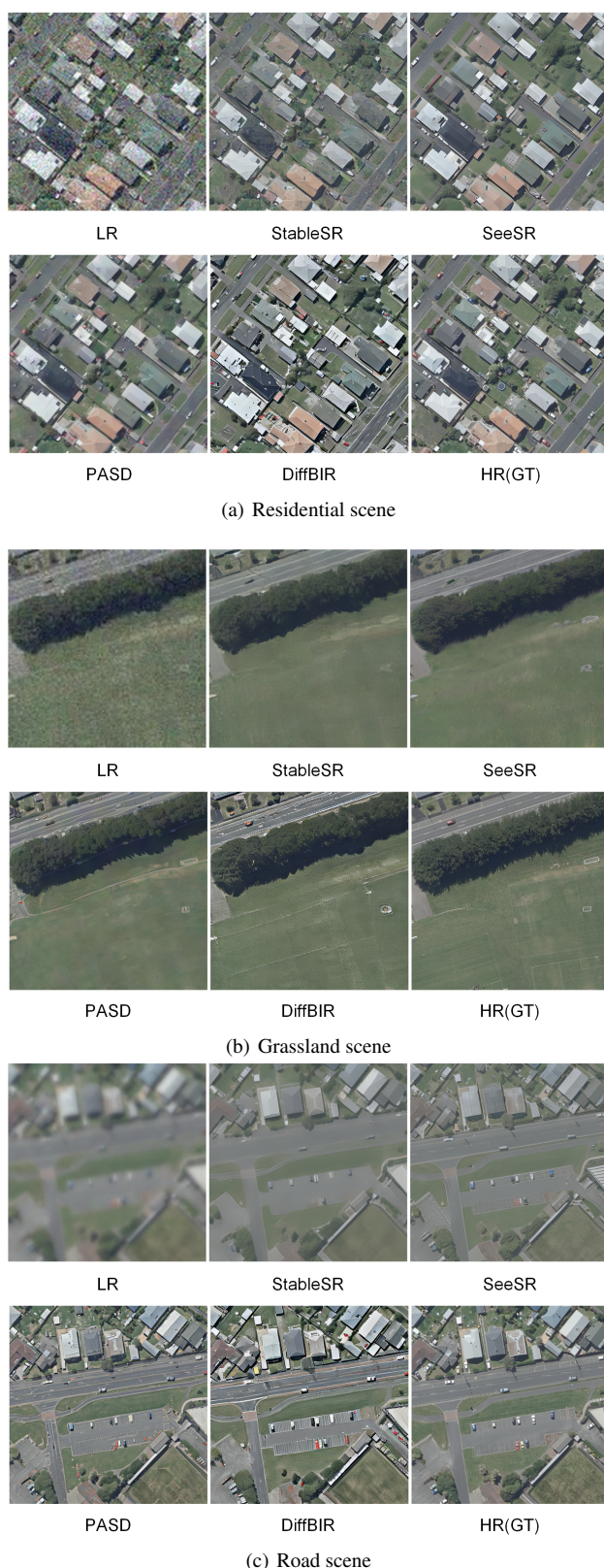


Figure 7. Qualitative comparison of SD-based super-resolution methods on three representative remote sensing scenes. From left to right in each row are low-resolution (LR), StableSR, SeeSR, PASD, DiffBIR, and high-resolution (HR) ground truth.

Table 3. Perceptual and no-reference quantitative comparison on the WHU-Mix test set. Best results are in bold and second-best results are underlined.

Method	FID↓	CLIP-IQA↑	MUSIQ↑	MANIQA↑
StableSR	156.47	0.34	37.78	0.21
SeeSR	128.26	<u>0.60</u>	64.57	0.46
PASD	<u>120.32</u>	0.34	43.45	0.26
DiffBIR	110.58	0.68	61.17	<u>0.45</u>

is slightly stronger in distortion-oriented reconstruction, while PASD provides a better balance between structural consistency and perceptual similarity. PASD also improves FID to 120.32, suggesting that its gains are not limited to full-reference metrics.

SeeSR shows a different profile. Although its PSNR and SSIM are lower than those of StableSR and PASD, it achieves the best MUSIQ (64.57) and MANIQA (0.46), together with the second-best CLIP-IQA (0.60) and LPIPS (0.47). This result indicates that semantics-aware guidance is particularly effective for improving perceptual quality and semantic plausibility.

DiffBIR is strongest in generative realism. It achieves the best FID of 110.58 and the best CLIP-IQA of 0.68, while remaining competitive on MUSIQ and MANIQA. However, this advantage is accompanied by the lowest PSNR (19.20 dB), the lowest SSIM (0.68), and the highest LPIPS (0.54), showing a clearer trade-off between realism and strict reference fidelity.

Overall, the four methods correspond to different priorities. StableSR is distortion-oriented, PASD is the most balanced in structure-preserving restoration, SeeSR is strongest in semantic and perceptual quality, and DiffBIR is the most realism-oriented. This differentiated behavior explains why no single method dominates across all metrics and suggests that method selection should depend on the target application.

3.6 Qualitative Comparison

Quantitative metrics alone cannot fully capture restoration behavior in remote sensing scenes. Figure 7 therefore presents three representative cases, namely a dense residential region, a grassland scene, and a road-dominant scene.

In the residential scene, StableSR restores the main layout conservatively but remains somewhat smooth around roofs and narrow roads. SeeSR improves semantic separation among roofs, roads, and vegetation, although some contours are still softer than those of the reference. PASD preserves roof edges and parcel organization most faithfully, while DiffBIR produces the sharpest appearance but also the strongest over-enhancement.

In the grassland scene, StableSR maintains the global pattern but weakens fine field texture. SeeSR recovers more coherent texture and clearer weak linear traces, which agrees with its strong no-reference scores. PASD is more restrained but stable at the grassland-tree boundary, whereas DiffBIR produces the most vivid texture and contrast, although some details appear more aggressively synthesized.

In the road scene, StableSR recovers the main layout reliably but softens fine boundaries. SeeSR improves semantic clarity, while PASD gives the most faithful reconstruction of road edges, parking-lot contours, and local geometry. DiffBIR again

produces the strongest sharpness and contrast, but also shows the clearest tendency toward over-enhancement.

Overall, the qualitative comparison is consistent with the quantitative results: StableSR is conservative, PASD is the most structurally reliable, SeeSR is strong in perceptual-semantic quality, and DiffBIR is the most realism-oriented.

3.7 Overall Discussion

The combined results show that SD-based remote sensing super-resolution should be interpreted as a trade-off problem rather than a single-metric ranking problem. Methods with stronger conditional control, such as StableSR and PASD, tend to achieve better fidelity and structural consistency, whereas methods with greater generative freedom, such as SeeSR and DiffBIR, obtain stronger perceptual quality and realism. For remote sensing applications, this distinction is important because different tasks place different weights on geometry preservation, semantic plausibility, and visual enhancement.

4. Conclusion and Future Work

This study presents a comparative evaluation of StableSR, SeeSR, DiffBIR, and PASD on the WHU-Mix remote sensing dataset. The results show that these methods do not follow a single ranking across all metrics: StableSR is strongest in PSNR (dB), PASD is best in SSIM and LPIPS while remaining competitive in FID, SeeSR is strongest in semantic and no-reference perceptual quality, and DiffBIR achieves the best FID and CLIP-IQA at the cost of weaker full-reference fidelity. These findings indicate that method selection should depend on application requirements rather than a single metric. At the same time, the study is limited to one dataset, one degradation family, and image-quality-oriented evaluation. Future work should therefore expand validation to more diverse remote sensing data, examine downstream task performance, and explore domain-adaptive diffusion models that better reconcile fidelity, structure, semantics, and realism.

References

- Hore, A., Ziou, D., 2010. Image Quality Metrics: PSNR vs. SSIM. *2010 20th International Conference on Pattern Recognition (ICPR)*, IEEE, 2366–2369.
- Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F., 2021. MUIQ: Multi-Scale Image Quality Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.
- Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C., 2024. DiffBIR: Toward Blind Image Restoration With Generative Diffusion Prior. *European Conference on Computer Vision*, 430–448.
- Luo, M., Ji, S., Wei, S., 2022. A Diverse Large-Scale Building Dataset and a Novel Plug-and-Play Domain Generalization Method for Building Extraction. *arXiv Preprint arXiv:2208.10004*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.
- Saxena, D., Cao, J., 2021. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3), 1–42.
- Wang, J., Chan, K. C. K., Loy, C. C., 2023. Exploring CLIP for Assessing the Look and Feel of Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2555–2563.
- Wang, J., Yue, Z., Zhou, S., Chan, K. C., Loy, C. C., 2024. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *International Journal of Computer Vision*, 5929–5949.
- Wang, X., Xie, L., Dong, C., Shan, Y., 2021. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1905–1914.
- Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 600–612.
- Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L., 2024. SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25456–25467.
- Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y., 2022. MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1191–1200.
- Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L., 2024. Pixel-Aware Stable Diffusion for Realistic Image Super-Resolution and Personalized Stylization. *European Conference on Computer Vision*, 74–91.
- Zhang, K., Liang, J., Van Gool, L., Timofte, R., 2021. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4791–4800.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.