FiDeSR: High-Fidelity and Detail-Preserving One-Step Diffusion Super-Resolution

Aro Kim[1*] Myeongjin Jang[1*] Chaewon Moon[1] Youngjin Shin[1] Jinwoo Jeong[2] Sang-hyo Park[1†]
[1]Kyungpook National University [2]Korea Electronics Technology Institute
{arokim37, jmj4431, moonchaewon2, 2021112241, s.park}@knu.ac.kr jw.jeong@keti.re.kr

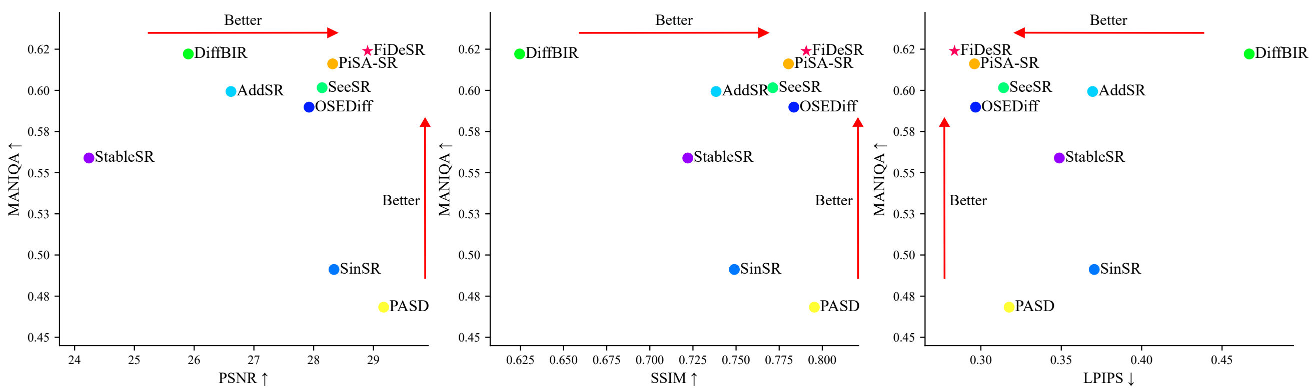

Figure 1. Performance comparison among Real-ISR methods on three perceptual–fidelity metric pairs: PSNR vs. MANIQA (left), SSIM vs. MANIQA (middle), and LPIPS vs. MANIQA (right). Higher MANIQA, PSNR, and SSIM values and lower LPIPS values indicate better performance. FiDeSR achieves superior perceptual quality while maintaining competitive fidelity across all three metric pairs. All methods are evaluated on the DRealSR dataset.

Abstract

Diffusion-based approaches have recently driven remarkable progress in real-world image super-resolution (SR). However, existing methods still struggle to simultaneously preserve fine details and ensure high-fidelity reconstruction, often resulting in suboptimal visual quality. In this paper, we propose FiDeSR, a high-fidelity and detail-preserving one-step diffusion super-resolution framework. During training, we introduce a detail-aware weighting strategy that adaptively emphasizes regions where the model exhibits higher prediction errors. During inference, low- and high-frequency adaptive enhancers further refine the reconstruction without requiring model retraining, enabling flexible enhancement control. To further improve the reconstruction accuracy, FiDeSR incorporates a latent residual refinement, which corrects prediction errors in the diffusion noise and enhances fine detail recovery. FiDeSR achieves superior real-world SR performance compared to existing diffusion-based methods, producing outputs with both high perceptual quality and faithful content restoration. The source code will be released at: `https://github.com/Ar0Kim/FiDeSR`.

*Equal contribution. †Corresponding author.

1. Introduction

Image super-resolution (ISR) is the task of restoring a high-quality (HQ) image from its low-quality (LQ) input. Traditional ISR methods aim to restore downsampled LQ images using simple bicubic interpolation [54]. However, images captured in real-world environments suffer from unknown and complex degradations. To address these degradations, real-world image super-resolution (Real-ISR) has been developed to restore natural and realistic HQ images.

Generative models have been utilized as a solution for producing realistic results in Real-ISR. Generative adversarial networks (GANs) [10, 20, 44] have shown success in generating realistic results. Recently, Diffusion Models (DMs) [13, 36] have emerged as a solution for Real-ISR [31], as their powerful image generation capability [6] enables the restoration of realistic images under various degradation conditions. However, diffusion-based Real-ISR methods require iterative sampling steps, leading to

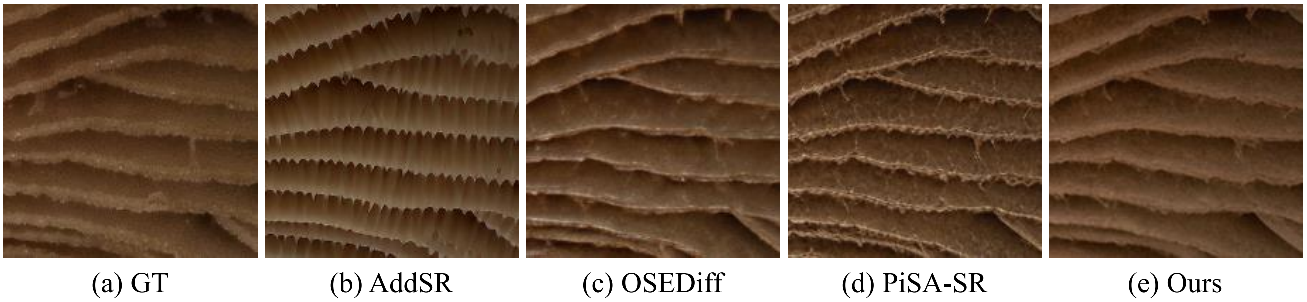

Figure 2. Example failure cases of diffusion-based Real-ISR methods. (b) AddSR introduces structural distortion and low-frequency inconsistency. (c) OSEDiff loses high-frequency details, producing over-smoothed texture. (d) PiSA-SR generates excessive details. In contrast, (e) our method achieves both high fidelity and detail-preserving.

high computational costs and long inference times [58]. To address this issue, efficient one-step diffusion super-resolution models [48, 52] have been proposed, which distill the multi-step diffusion process into a one-step sampling framework while maintaining reconstruction quality [57].

Among one-step diffusion-based Real-ISR methods, there remain two challenging problems: **(1) High-fidelity, (2) Detail restoration**. These approaches often suffer from structural distortion and low-frequency (LF) inconsistency due to VAE-based conditioning, making it difficult to preserve faithful image content while recovering fine high-frequency details, as observed in Fig. 2(b) [2, 38]. In addition, one-step diffusion models also struggle with high-frequency (HF) detail loss. Diffusion models generate HF details to compensate for the loss of HF information, which is degraded during the noise injection process of diffusion. In particular, latent diffusion models (LDMs) further amplify this loss, as the variational autoencoder (VAE) encoder compresses pixel-level representations into a compact latent space for efficient computation [30]. While multi-step diffusion [43] enables sufficient or excessive detail restoration by generating HF details through iterative denoising, one-step diffusion reduces this process to a single step, resulting in insufficient HF detail restoration [8, 52], as illustrated in Fig. 2(c). Additionally, recent one-step diffusion models [8, 39] aim to predict the residual between the LQ and HQ latent representations. This residual learning strategy simplifies the SR process by allowing the model to focus on recovering HF information within the latent, leading to faster convergence and more efficient training. However, since the U-Net predicts only a single global residual at each stage, this approach results in unstable HF reconstruction and residual artifacts, as shown in Fig. 2(d).

To overcome these limitations, we propose **a high-fidelity and detail-preserving one-step diffusion-based super-resolution (FiDeSR)** framework, which introduces three key components. First, we employ a **Detail-aware Weighting (DAW)** strategy that utilizes frequency-based information to emphasize important detail-rich regions during training. DAW adaptively emphasizes regions where the current diffusion-based one-step model underperforms, rather than allowing it to overfit to already well-reconstructed areas. Second, we incorporate a **Latent Residual Refinement Block (LRRB)** to compensate for incomplete HF reconstruction and artifacts arising from coarse residual prediction. LRRB enhances the representational capacity of the model, while maintaining the efficiency of one-step diffusion. Finally, to simultaneously enhance perceptual detail and maintain fidelity, we introduce a **Latent Frequency Injection Module (LFIM)** that selectively injects frequency components extracted from the denoised latent according to spatial and channel characteristics.

As shown in Fig. 1, FiDeSR achieves superior perceptual quality while maintaining competitive fidelity on real-world super-resolution benchmarks, thereby demonstrating its ability to reconstruct fine details while preserving high structural fidelity compared to state-of-the-art diffusion-based Real-ISR methods. Our main contributions are summarized as follows:

- We propose FiDeSR, a high-fidelity and detail-preserving one-step diffusion-based Real-ISR framework that effectively addresses challenges of structural fidelity degradation and insufficient high-frequency restoration in one-step diffusion models.
- We introduce three key technical components to form the FiDeSR framework: the Detail-aware Weighting (DAW) strategy, the Latent Residual Refinement Block (LRRB), and the Latent Frequency Injection Module (LFIM). These components are specifically designed to collectively address the challenges of high-fidelity and detail restoration in one-step diffusion-based Super-Resolution.
- FiDeSR achieves superior detail reconstruction and structural consistency on real-world SR benchmarks, outperforming state-of-the-art one-step and even competitive multi-step diffusion Real-ISR methods.

2. Related Works

2.1. Multi-step Diffusion-based Super-Resolution

Diffusion models have been widely adopted in Real-ISR based on their strong generative priors. Early works applied pixel-space diffusion models such as DDPM [13] for iterative refinement [21, 29, 31, 33, 58], but these approaches suffer from high computational cost. To improve efficiency, later methods operate in the latent space of pretrained T2I diffusion models like Stable Diffusion [30], often combined with PEFT techniques such as LoRA [14, 15]. Representative approaches include StableSR [43], DiffBIR [24], PASD [56], and SeeSR [53], which leverage semantic prompts, feature injection, or restoration priors to guide the diffusion process. However, these methods still rely on multi-step denoising, resulting in slow inference.

2.2. One-step Diffusion-based Super-Resolution

To accelerate inference, recent studies compress the iterative diffusion process into a single forward pass via distillation. SinSR [48] adopts deterministic sampling and consistency-preserving loss to emulate multi-step behavior. AddSR [40] applies adversarial diffusion distillation [32] to reduce sampling steps, while OSEDiff [52] uses VSD [50] with LoRA-based fine-tuning to restore images directly from LQ latents. PiSA-SR [39] introduces Dual-LoRA for controllable pixel and semantic refinement, and TSD-SR [8] improves VSD stability using Target Score Distillation within SD3 [9]. While TSD-SR performs well in terms of quantitative metrics, it exhibits noticeable grid artifacts, a limitation inherited from its SD3-based architecture. Although highly efficient, these one-step approaches still struggle with insufficient high-frequency detail, misalignment between LF structures and generated HF textures, and limitations of predicting a single global residual [8, 39].

2.3. Frequency-Domain Approaches for Super-Resolution

Frequency-domain information has long been leveraged in SR to recover high-frequency details lost in degraded inputs [4, 5, 11, 16, 19, 28]. Recent diffusion-based SR methods have incorporated frequency cues to enhance generative restoration. RnG [47] separates LF reconstruction and HF residual generation but relies on computationally expensive multi-step residual diffusion with a patch-wise step controller, which increases complexity and limits scalability. DiWa [27] and Frequency-Domain Refinement [46] perform diffusion directly in the wavelet or frequency domain for finer HF recovery. FreeU [35] modulates diffusion features to strengthen HF reconstruction without retraining. Multi-Scale Generation Guidance [34] combines GAN supervision with DWT-based HF losses but still depends on iterative sampling and adversarial training. GuideSR [2] improves structural fidelity through full-resolution guidance in a one-step framework, yet remains biased toward fidelity and less effective in perceptual realism. TFDSR [23] incorporates frequency information into multi-step diffusion models by adapting frequency enhancement across diffusion timesteps within the denoising process. Despite these advances, most prior methods emphasize either high- or low-frequency components in isolation, or require additional overhead when addressing both. Thus, there is a need for a one-step diffusion framework that can effectively account for both frequency components and restore fine details and overall fidelity in a balanced manner.

3. Method

3.1. Preliminaries

The objective of Real-ISR is to restore a corresponding HQ image x_H from a counterpart x_L that has a complex and unknown degradation process. The Real-ISR task is formulated as an optimization problem for a restoration network G_θ with parameters θ, using (x_L, x_H) pairs sampled from the training data distribution D. The objective is to find the optimal parameters θ^* that minimize the following composite function:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x_L,x_H)\sim D} \left[\mathcal{L}_{rec}(x_{SR}, x_H) + \mathcal{L}_{reg}(x_{SR})\right]. \tag{1}$$

This objective function attempts to balance two goals to achieve realistic restoration results. $\mathcal{L}_{rec}$ is the reconstruction loss, which enforces that the restored image x_{SR} remains faithful to the HQ image x_H, typically measured using pixel-wise losses or perceptual metrics. This term is balanced by $\mathcal{L}_{reg}$, regularization loss, which guides the restored image x_{SR} to follow the statistical properties of natural images. How to define and optimize the $\mathcal{L}_{reg}$ term in Eq. 1 has become the core challenge in Real-ISR.

Recently, Real-ISR methods leverage diffusion-based generative priors to satisfy this $\mathcal{L}_{reg}$ for more realistic restoration. To reduce the high computational cost of multi-step diffusion inference, one-step diffusion SR methods compress the iterative denoising into a single step. In particular, PiSA-SR [39] performs restoration directly in the latent space of a pretrained VAE, where a U-Net predicts a global residual to bridge the latent representations between the LQ and HQ images in a single step. This residual-based restoration can be generally formulated as:

$$z_0 = z_L - r, \tag{2}$$

where z_L denotes the latent representation of the LQ image, r is the residual predicted by the network, and z_0 denotes the restored latent obtained by applying the predicted residual to z_L.

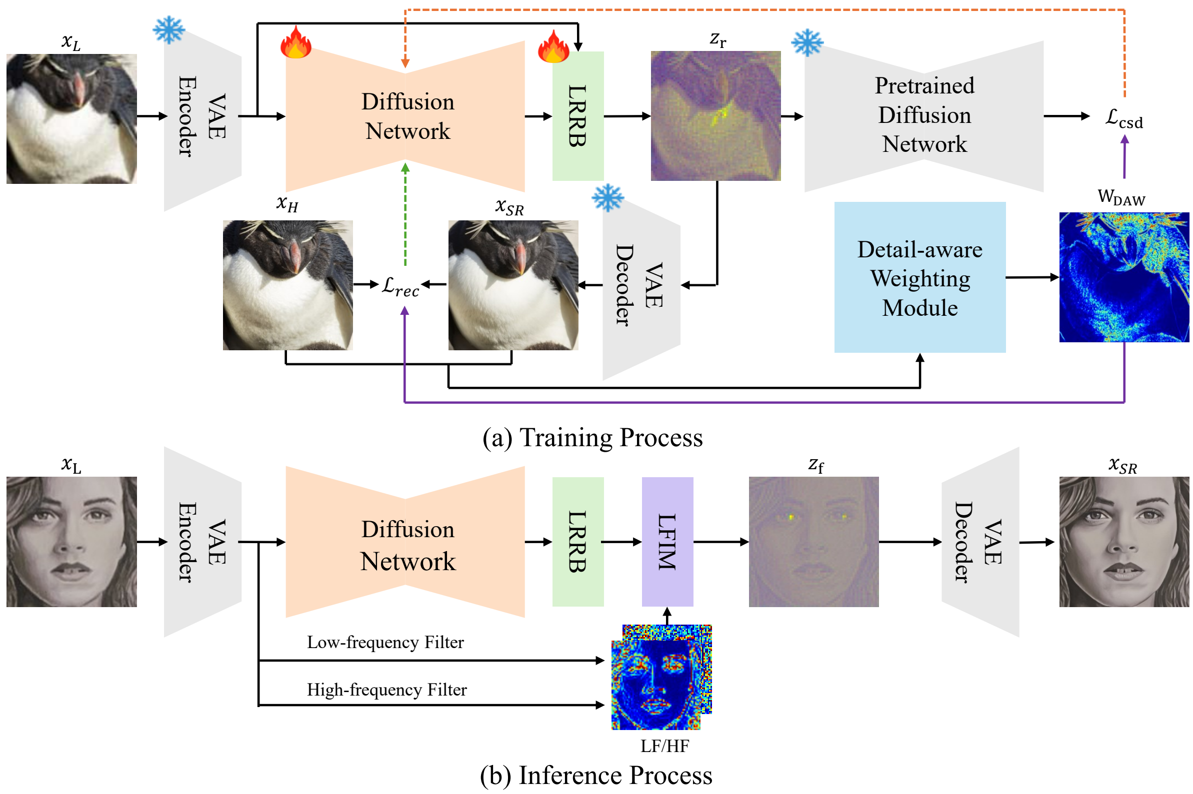

Figure 3. Overall framework of FiDeSR. (a) Training process: LQ image x_L is encoded into a latent z_L, and the Diffusion Network predicts a coarse residual r, which is refined by LRRB to create refined latent z_r. The training loss is guided by the DAW module to emphasize fine structural details, and the refined latent z_r is decoded by the VAE Decoder. (b) Inference process: Following the single-step diffusion process, the LQ latent z_L is processed by the Diffusion Network to predict a residual, which is then refined by the LRRB. Refined z_r is enhanced by frequency components through the LFIM, and finally decoded by the VAE Decoder to produce the Super-Resolution (SR) image x_{SR}.

3.2. Overview of FiDeSR

As shown in Fig. 3(a), our FiDeSR framework aims to achieve high-fidelity, detail-preserving one-step super-resolution by integrating Detail-aware Weighting (DAW) and latent residual refinement block (LRRB) into the one-step diffusion process. Given a LQ input image x_L, we first encode it into the latent representation z_L using a pretrained VAE encoder. U-Net student network G_θ predicts a coarse residual r to approximate the degradation between z_L and its HQ counterpart z_H. This initial residual is further refined by our LRRB, which learns an adaptive correction Δr based on both z_L and r. The refined residual r' is then used to reconstruct the refined latent z_r, which is decoded by the VAE decoder to produce the restored image. During training, FiDeSR employs a DAW module to adaptively weight losses based on a spatial detail map computed from Sobel, Laplacian, and Variance filters. By emphasizing regions rich in edges and textures, DAW guides the model to focus on fine structural recovery and visually important details, leading to sharper and more faithful restoration.

3.3. Detail-aware Weighting

To adaptively emphasize regions that contain rich structural details and perceptually important textures, we introduce the DAW module. Unlike conventional frequency-based methods that explicitly decompose images in the Fourier domain, our DAW dynamically adjusts the loss by directly extracting HF details in the spatial domain. We construct a detail map D to guide this weighting process. Specifically, multiple spatial operators including Sobel, Laplacian, and local variance filters are applied to the HQ image x_H to capture edge sharpness, local contrast, and texture variance. The detail map is defined as the mean response of these operators:

$$D = \frac{Sobel(x_H) + Laplacian(x_H) + Variance(x_H)}{3}. \tag{3}$$

In addition, we compute an error map E between the restored image x_{SR} and the ground-truth image x_H, which captures both pixel-level and perceptual discrepancies. The pixel-wise error E_{pix} is computed using the L1 difference,

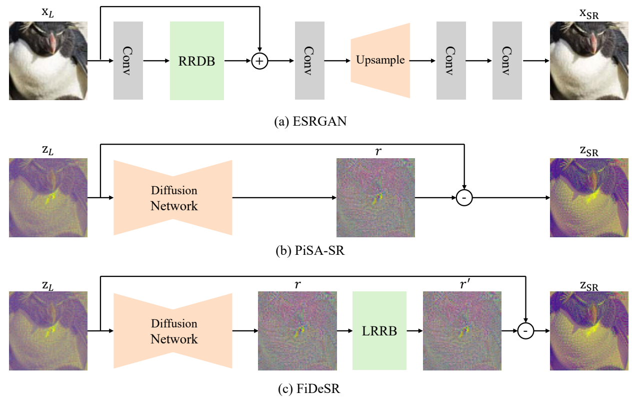

Figure 4. Comparison of the pipeline among ESRGAN, PiSA-SR, and our FiDeSR. (a) ESRGAN employs RRDB-based restoration in pixel space. (b) PiSA-SR uses a one-step diffusion network to predict a global residual in the latent space. (c) Our FiDeSR introduces latent residual refinement block (LRRB) that refines the residual.

and the perceptual error E_{perc} is computed using LPIPS. The pixel-wise error map E_{pix} is defined as:

$$E_{pix} = |x_{SR} - x_H|. \tag{4}$$

The perceptual error map E_{perc} is obtained using an LPIPS network:

$$E_{perc} = LPIPS(x_{SR}, x_H). \tag{5}$$

The final error map is computed as a weighted combination of pixel-level and perceptual errors:

$$E = (1 - p)\, E_{pix} + p\, E_{perc}, \tag{6}$$

where p controls the contribution of perceptual discrepancy.

The DAW module integrates these two components to produce a difficulty weight map W_{DAW}, which represents the spatial difficulty of each pixel. This map is computed by element-wise multiplying the detail map D and the error map E:

$$W_{DAW} = D \odot E. \tag{7}$$

This map adaptively scales the loss contribution of each spatial position, assigning higher weights to visually complex or detail-rich regions. We apply this weighting strategy to both the reconstruction loss and the classifier score distillation (CSD) loss [39, 50], enabling the model to emphasize semantically and structurally difficult regions across all spatial objectives. By applying this strategy, our method effectively focuses the difficulty-aware weighting on detail-rich regions, which guides the model to synthesize more accurate HF details, resulting in sharper textures, improved structural consistency, and enhanced visual realism.

3.4. Latent Residual Refinement Block

Existing one-step diffusion-based super-resolution methods (e.g., PiSA-SR [39]) lack an iterative correction process, often resulting in unstable latent residual prediction and insufficient high-frequency recovery. To address this limitation, we introduce the LRRB, as shown in Fig. 4(c).

LRRB operates in the latent space to refine the residual predicted by the diffusion U-Net before it is subtracted from the LQ latent z_L. Built upon the Residual-in-Residual Dense Block (RRDB) [44], LRRB differs from pixel-domain residual refinement (e.g., ESRGAN), which assumes stable residual prediction. Instead, it explicitly targets diffusion-induced instability in latent residual prediction.

The LRRB takes the concatenation of the LQ latent z_L and the initial residual of the U-Net r as input. This combined latent is first mapped to an intermediate latent space by an initial 1x1 convolution, then passes through several dense block modules. Finally, a 1x1 convolution remaps these latents to the original residual channel dimension, generating the predicted correction value Δr.

This predicted correction Δr is added to the original residual of the U-Net r to create a refined residual r':

$$r' = r + \Delta r. \tag{8}$$

This refined residual r' is ultimately subtracted from the LQ latent z_L to compute the refined latent z_r:

$$z_r = z_L - r'. \tag{9}$$

This refined latent z_r is then passed to the VAE decoder to generate the final SR image.

This LRRB effectively advances the conventional simple residual subtraction method, as shown in Eq. 2, into a more powerful learning-based refinement stage. Instead of relying on U-Net initial prediction r, this output is treated as a strong initial estimate. LRRB then learns to predict an optimal correction Δr based on both r and the context of z_L. This two-step process, shown in Eq. 8 and 9, provides the model with greater flexibility to make more precise and complex adjustments to the residual. This learned refinement ultimately leads to a more accurate refined latent z_r and a corresponding improvement in reconstruction quality.

3.5. Training Losses

The FiDeSR framework is optimized with a composite objective that enforces both pixel-level fidelity and semantic consistency. The total training loss consists of two key components: the reconstruction loss $\mathcal{L}_{rec}$ and the regularization loss $\mathcal{L}_{reg}$. Each of these loss terms is spatially modulated by the difficulty weight map W_{DAW}. This weighting guides the network to focus more on edges, textures, and fine details, resulting in sharper and more realistic reconstructions.

The reconstruction loss $\mathcal{L}_{rec}$ is formulated as a sum of the pixel-wise MSE loss and the perceptual LPIPS loss, both spatially weighted by our W_{DAW} map. To account for the different resolutions, the MSE loss is weighted by W_{DAW} at the original image resolution, while the LPIPS loss is weighted by W'_{DAW}, which is an interpolated version of W_{DAW} to match the LPIPS feature map dimensions:

$$\begin{aligned}\mathcal{L}_{rec} = \lambda_{mse}\, \mathbb{E}\big[W_{DAW} \cdot (x_{SR} - x_H)^2\big] \\ + \lambda_{lpips}\, \mathbb{E}[W'_{DAW} \cdot \mathcal{L}_{LPIPS\_map}]\,.\end{aligned} \tag{10}$$

The regularization term $\mathcal{L}_{reg}$ is implemented as a CSD loss, which distills the semantic priors from a pretrained diffusion model to guide restored images towards natural and semantically consistent distributions with improved stability and efficiency. We modified the $\mathcal{L}_{CSD}$ function to accept the W_{DAW} map as an argument. This allows the module to internally apply its semantic guidance with a focus on the critical regions identified by the map:

$$\mathcal{L}_{reg} = \lambda_{reg} \cdot \mathcal{L}_{CSD}(z_r, prompt, W_{DAW}). \tag{11}$$

The total training objective $\mathcal{L}_{total}$ combines these components, following the structure established in Eq. 1:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{reg}. \tag{12}$$

The model is guided by W_{DAW} to synthesize more accurate HF details, resulting in sharper textures, improved structural consistency, and enhanced visual realism.

3.6. LF/HF Adaptive Enhancers for Inference

As illustrated in Fig. 3(b), the inference process of FiDeSR restores a LQ input x_L into a SR output x_{SR} by combining the diffusion process with frequency-based detail enhancements. Following the diffusion process, the latent feature z_L encoded by the VAE Encoder performs single-step restoration through the U-Net. The initial residual ($\mathbf{r}$) predicted by the U-Net is refined through the LRRB, resulting in a refined residual r'. This refined residual is then subtracted from z_L to compute the refined latent z_r.

The refined latent z_r is decomposed into low-frequency $\boldsymbol{\Delta}_{LP}$ and high-frequency $\boldsymbol{\Delta}_{HP}$ components by FFT-based Butterworth filters. These frequency components are selectively injected into z_r through the latent frequency injection module (LFIM). The LFIM consists of a spatial gate M_{sp}, which identifies detailed and flat regions based on a detail map (Sobel, Laplacian, Variance) derived from the LQ image x_L, and a channel gate M_{ch}, which analyzes the frequency energy ratio of each latent channel. This selective injection strategy focuses on low-frequency enhancement in structure and high-frequency enhancement in texture, maximizing detail recovery. The final enhanced latent z_f is passed to the VAE Decoder to generate the final HQ image x_{SR} in pixel space. More details are provided in the supplementary material.

4. Experiments

4.1. Experimental Settings

Training Datasets. We employed LSDIR [22], DIV2K [1], Flickr2K [41], and the first 10K images from FFHQ [17] as training data. We generated LQ-HQ pairs for Real-ISR training by degrading HR images using the Real-ESRGAN [45] degradation pipeline.

Test Datasets. We evaluate our method on both synthetic and real-world datasets. The synthetic test dataset consists of 3,000 images of 512x512, cropped from DIV2K-validation [1] and degraded into 128x128 LQ images using the Real-ESRGAN degradation pipeline. The real-world datasets are obtained by center-cropping images from RealSR [3] and DRealSR [51], where the LQ and HQ images have resolutions of 128x128 and 512x512.

Evaluation Metrics. To comprehensively evaluate the performance of our method, we employ both full-reference and no-reference metrics. For full-reference metrics, PSNR and SSIM [49] are used to measure pixel-level fidelity between restored and GT images. LPIPS [59] and DISTS [7] are adopted to evaluate perceptual quality by comparing deep feature similarities. FID [12] is utilized to evaluate the distributional distance between the restored and GT image sets. No-reference metrics are used to evaluate perceptual quality without relying on GT images, including CLIPIQA [42], NIQE [26], MUSIQ [18], and MANIQA [55].

Implementation details. Our method is implemented based on the Stable Diffusion 2.1-base [37]. The pretrained VAE and U-Net are kept frozen, and we add trainable LoRA layers to the U-Net for fine-tuning, setting the LoRA rank to 8. We train our model on 2 NVIDIA H100 GPUs with a batch size of 8 for 200K training steps. We used the AdamW optimizer [25] with a learning rate of 5×10^{-5}. We utilize RAM [60] to extract text prompts. In all experiments, the loss weights are set to $\lambda_{\text{mse}} = 1$ and $\lambda_{\text{lpips}} = 2$.

4.2. Comparison with existing methods

Quantitative comparisons. Table 1 provides a quantitative comparison with recent diffusion-based SR methods on synthetic and real-world benchmarks. Across all datasets, FiDeSR shows consistently reliable performance, maintaining a balanced improvement in both full-reference and no-reference metrics while requiring only a single diffusion step. In particular, FiDeSR achieves consistently better scores in perceptual similarity measures such as LPIPS and DISTS, indicating improved reconstruction and reduced artifacts compared to other one-step and multi-step diffusion models. At the same time, FiDeSR also demonstrates competitive no-reference performance on metrics such as CLIP-IQA, MUSIQ, and MANIQA, suggesting that the restored images better align with natural image statistics and exhibit more coherent visual structure. This si-

Table 1. Quantitative comparison with state-of-the-art DM-based SR methods on synthetic and real-world test datasets. The number of diffusion inference steps is indicated by 's'. The best and second best results of each metric are highlighted in **red** and **blue**, respectively. The third best results are indicated by an **underline**.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	CLIPIQA ↑	NIQE ↓	MUSIQ ↑	MANIQA ↑	FID ↓
DRealSR	StableSR-200s	27.93	0.7491	0.3306	0.2290	0.6197	6.3865	58.5502	0.5569	147.48
	SeeSR-50s	28.14	0.7713	0.3141	0.2298	0.6889	6.4633	64.7302	0.6016	146.98
	DiffBIR-50s	25.90	0.6245	0.4669	0.2298	0.6333	**6.3275**	**66.1334**	**0.6221**	180.34
	PASD-20s	**29.17**	**0.7954**	0.3176	0.2220	**0.7071**	7.7138	50.4422	0.4683	143.08
	AddSR-4s	26.62	0.7384	0.3695	0.2651	**0.7121**	7.8536	65.0625	0.5993	167.69
	OSEDiff-1s	27.92	**0.7835**	**0.2967**	**0.2162**	0.6956	6.4389	64.6979	0.5898	**135.45**
	SinSR-1s	**28.34**	0.7490	0.3705	0.2487	0.6445	7.0300	55.3668	0.4912	175.96
	PiSA-SR-1s	28.32	0.7804	**0.2960**	**0.2169**	0.6918	**6.1803**	**66.1112**	**0.6161**	**130.48**
	FiDeSR-1s	**28.90**	**0.7907**	**0.2836**	**0.2112**	**0.6974**	**6.2014**	**65.7820**	**0.6239**	**127.97**
RealSR	StableSR-200s	24.73	0.7084	0.3050	0.2156	0.6379	**5.6251**	65.4419	0.6259	130.43
	SeeSR-50s	25.21	0.7216	0.3004	0.2218	0.6674	5.5929	69.6929	0.6435	125.09
	DiffBIR-50s	24.83	0.6501	0.3650	0.2399	**0.7053**	5.8404	69.2781	0.6502	130.52
	PASD-20s	**26.63**	**0.7660**	**0.2869**	**0.2032**	0.4857	5.9815	60.0287	0.5390	124.52
	AddSR-4s	22.54	0.6419	0.3867	0.2718	**0.7298**	6.5640	**71.4404**	**0.6751**	156.68
	OSEDiff-1s	25.15	0.7341	0.3194	0.2127	0.6686	5.6364	69.0810	0.6335	**123.49**
	SinSR-1s	**26.29**	0.7350	0.3194	0.2344	0.6181	6.2854	60.7494	0.5420	134.27
	PiSA-SR-1s	25.50	**0.7418**	**0.2672**	**0.2044**	0.6697	**5.5054**	**70.1462**	**0.6551**	**124.18**
	FiDeSR-1s	**26.02**	**0.7457**	**0.2626**	**0.1965**	**0.6896**	**5.3194**	**69.8245**	**0.6681**	**109.68**
DIV2K	StableSR-200s	23.27	0.5733	0.3106	0.2045	0.6773	4.7621	65.8259	0.6178	**24.43**
	SeeSR-50s	23.73	0.6057	0.3193	**0.1966**	0.6864	4.7949	68.3976	0.6200	25.80
	DiffBIR-50s	23.14	0.5441	0.3669	0.2209	**0.7299**	4.9920	**69.8698**	**0.6440**	32.71
	PASD-20s	**24.41**	**0.6252**	0.3794	0.2218	0.5567	5.4273	61.2676	0.5362	31.53
	AddSR-4s	22.38	0.5554	0.3815	0.2340	**0.7538**	5.8168	**69.1623**	0.6300	35.07
	OSEDiff-1s	23.72	**0.6109**	**0.2942**	0.1975	0.6681	**4.7104**	67.9670	0.6131	26.33
	SinSR-1s	**24.40**	0.6018	0.3244	0.2067	0.6492	6.0002	62.8673	0.5393	35.59
	PiSA-SR-1s	23.87	0.6058	**0.2823**	**0.1934**	**0.6928**	**4.5563**	**69.6799**	**0.6375**	**25.09**
	FiDeSR-1s	**24.33**	**0.6250**	**0.2678**	**0.1845**	0.6873	**4.6644**	68.8672	**0.6384**	**23.30**

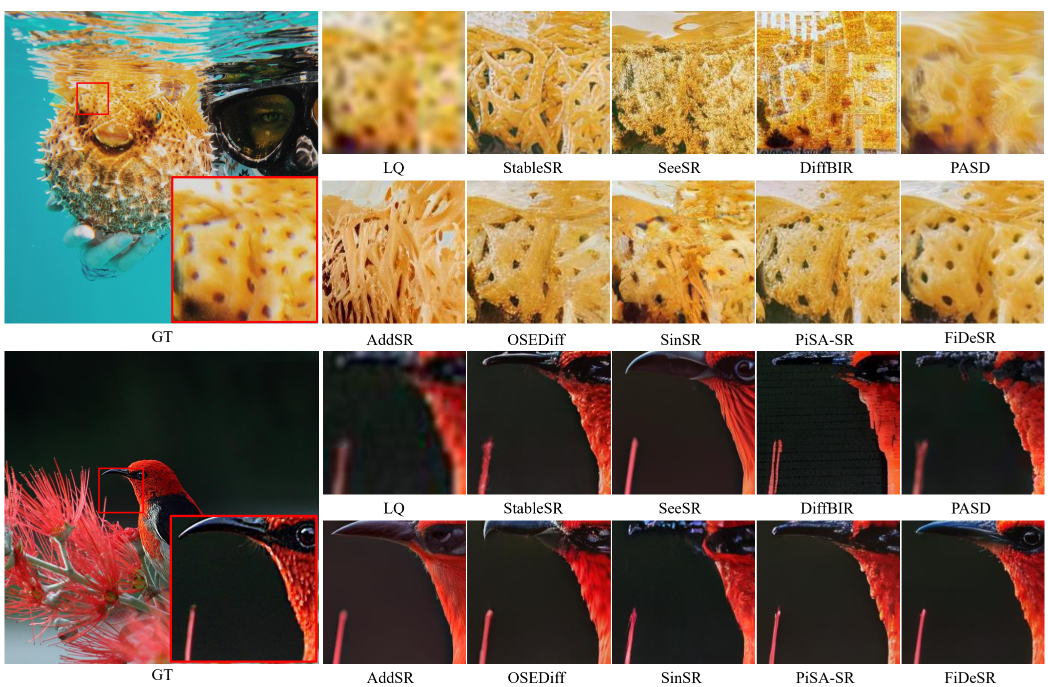

Figure 5. Qualitative comparisons with state-of-the-art DM-based SR methods.

Table 2. Ablation study on LRRB and DAW on the DIV2K dataset (evaluated before applying LFIM).

Model	CLIPIQA ↑	NIQE ↓	MUSIQ ↑	MANIQA ↑
w/o LRRB, DAW	0.6611	4.7381	67.6043	0.6237
w/o LRRB	0.6641	4.7129	67.6280	0.6236
w/o DAW	0.6626	4.7340	67.9515	0.6278
FiDeSR	**0.6699**	**4.6300**	**68.2869**	**0.6285**

Table 3. High-frequency noise prediction error comparison between baseline and LRRB models. Lower MSE indicates better performance.

Method	DIV2K	DRealSR	RealSR	Average
baseline	0.1051	0.1049	0.1045	0.1048
LRRB	0.1038	0.1028	0.1029	0.1032
Improvement	1.24%	2.00%	1.53%	1.59%

multaneous improvement on both reference-based and no-reference evaluation is challenging due to the inherent perception–distortion trade-off, yet FiDeSR manages to maintain a more stable balance than prior diffusion-based SR approaches. Furthermore, FiDeSR achieves consistently lower FID values than all existing methods including both one-step and multi-step methods, demonstrating the closest alignment with real-image distributions overall. Overall, FiDeSR demonstrates consistently superior restoration performance across diverse real-world settings by effectively addressing the common limitations of one-step diffusion, overcoming insufficient high-frequency detail reconstruction and low-frequency structural inconsistency.

Qualitative comparisons. Visual results are shown in Fig. 5. Multi-step methods such as SeeSR generate rich details, but many of them appear unnatural, often introducing noise-like artifacts and compromising fidelity. AddSR tends to deviate from the ground-truth structure, producing distorted shapes and lacking fine local details. DiffBIR also exhibits noticeable noise artifacts, while PASD often produces blurry textures and degraded structural consistency. OSEDiff preserves some coarse structures but still suffers from reduced fidelity and insufficient high-frequency reconstruction. PiSA-SR improves perceptual sharpness but may introduce overly enhanced or inconsistent textures in complex regions. In contrast, FiDeSR restores both structural integrity and fine details more faithfully to the ground truth while simultaneously producing sharper textures and a more natural overall appearance, without introducing unwanted artifacts.

4.3. Ablation study

Effectiveness of LRRB and DAW. Table 2 shows that both the LRRB and DAW modules contribute to perceptual quality. Removing either module consistently degrades the CLIPIQA, NIQE, MUSIQ, and MANIQA scores, indicating reduced detail preservation. When used together, FiDeSR achieves the best performance across all metrics,

Table 4. Ablation of LFIM (HF/LF) modules with different injection on the RealSR dataset.

Method	PSNR ↑	SSIM ↑	MUSIQ ↑	MANIQA ↑
Baseline	26.2542	0.7498	69.3610	0.6580
HF-0.1	26.1295	0.7476	69.6450	0.6639
HF-0.2	25.9985	0.7452	69.8562	0.6692
HF-0.3	25.8618	0.7428	70.0120	0.6737
HF-0.4	25.7213	0.7403	70.1132	0.6774
HF-0.5	25.5778	0.7378	70.1617	0.6803
LF-0.1	26.2905	0.7506	69.3177	0.6571
LF-0.2	26.3066	0.7513	69.2337	0.6564
LF-0.3	26.3221	0.7520	69.1431	0.6550
LF-0.4	26.3377	0.7526	69.0521	0.6538
LF-0.5	26.3448	0.7532	68.9473	0.6522

demonstrating the complementary effect of both modules in enhancing detail-aware restoration. The quantitative comparison in Table 3 further validates the benefit of incorporating the LRRB module. Across DIV2K, DRealSR, and RealSR datasets, the LRRB-equipped model consistently reduces the high-frequency noise prediction error compared to the baseline. More comprehensive results and analyses can be found in the supplementary materials.

Ablation on LFIM (HF/LF). Table 4 summarizes the effects of LFIM on both low- and high-frequency components. Applying LFIM on low-frequency features consistently improves PSNR and SSIM as the injection intensity increases, which confirms that reinforcing low-frequency information stabilizes global structure, tone, and illumination, thereby enhancing structural fidelity. In contrast, applying LFIM on high-frequency components substantially enhances perceptual quality metrics such as MUSIQ and MANIQA by selectively enhancing fine textures, edges, and high-frequency details. This demonstrates that the LFIM effectively complements structural refinement with perceptual sharpness, contributing to balanced and detail-aware super-resolution.

5. Conclusion

In this work, we presented FiDeSR that addresses two challenges of existing one-step diffusion methods: structural fidelity degradation and insufficient high-frequency detail restoration. To overcome these challenges, FiDeSR introduces three components: detail-aware weighting, latent residual refinement block, and the latent frequency injection module. Through extensive experiments, FiDeSR demonstrates state-of-the-art performance among one-step diffusion SR models and achieves consistent gains in both full-reference and no-reference metrics. Overall, FiDeSR highlights that one-step diffusion models can achieve high perceptual quality without sacrificing fidelity, when both frequency-aware guidance and residual refinement are properly integrated, which would open new directions for efficient real-world SR and extensions to video or multi-modal restoration tasks.

6. Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2025-00520308, 50%), in part by the Regional Innovation System & Education (RISE) Glocal 30 program through the Daegu RISE Center, funded by the Ministry of Education (MOE) and Daegu, Republic of Korea (2025-RISE-03-001, 25%), and in part by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4120242021487l, 25%).

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 6

[2] Aditya Arora, Zhengzhong Tu, Yufei Wang, Ruizheng Bai, Jian Wang, and Sizhuo Ma. Guidesr: Rethinking guidance for one-step high-fidelity diffusion-based super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 3973–3981, 2025. 2, 3

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3086–3095, 2019. 6

[4] Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1764–1774, 2023. 3

[5] Tao Dai, Jianping Wang, Hang Guo, Jinmin Li, Jinbao Wang, and Zexuan Zhu. Freqformer: Frequency-aware transformer for lightweight image super-resolution. In *Proceedings of the International joint conference on artificial intelligence*, pages 731–739, 2024. 3

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[7] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 6

[8] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23174–23184, 2025. 2, 3

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International conference on machine learning*, 2024. 3

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[11] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 104–113, 2017. 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 3

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *International conference on learning representations*, 1(2):3, 2022. 3

[16] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13919–13929, 2021. 3

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[18] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6

[19] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 3

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[21] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 3

[22] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1775–1787, 2023. 6

[23] Yueying Li, Hanbin Zhao, Jiaqing Zhou, Guozhi Xu, Tianlei Hu, Gang Chen, and Haobo Wang. A timestep-adaptive frequency-enhancement framework for diffusion-based image super-resolution. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1503–1511, 2025. 3

[24] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pages 430–448. Springer, 2024. 3

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[27] Brian B Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution. In *International joint conference on neural networks*, pages 1–8. IEEE, 2024. 3

[28] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021. 3

[29] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *European conference on computer vision*, pages 285–303. Springer, 2024. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[31] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 1, 3

[32] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European conference on computer vision*, pages 87–103. Springer, 2024. 3

[33] Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8975–8983, 2024. 3

[34] Yanli Shi, Xianhe Zhang, Yi Jia, and Jinxing Zhao. Multi-scale adversarial diffusion network for image super-resolution. *Scientific Reports*, 15(1):11690, 2025. 3

[35] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4733–4743, 2024. 3

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[37] Stability.ai. Stable diffusion. `https://stability.ai/stable-diffusion`. 6

[38] Lingchen Sun, Rongyuan Wu, Jie Liang, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability and efficiency of diffusion models for content consistent super-resolution. *arXiv preprint arXiv:2401.00877*, 2023. 2

[39] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2333–2343, 2025. 2, 3, 5

[40] Ying Tai, Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *Pattern recognition*, page 113012, 2026. 3

[41] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 6

[42] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 6

[43] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International journal of computer vision*, 132(12):5929–5949, 2024. 2, 3

[44] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision workshops*, pages 0–0, 2018. 1, 5

[45] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 6

[46] Xingjian Wang, Li Chai, and Jiming Chen. Frequency-domain refinement with multiscale diffusion for super resolution. *arXiv preprint arXiv:2405.10014*, 2024. 3

[47] Yujin Wang, Lingen Li, Tianfan Xue, and Jinwei Gu. Reconstruct-and-generate diffusion model for detail-preserving image denoising. *arXiv preprint arXiv:2309.10714*, 2023. 3

[48] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pages 25796–25805, 2024. 2, 3

[49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[50] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36: 8406–8441, 2023. 3, 5

[51] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European conference on computer vision*, pages 101–117. Springer, 2020. 6

[52] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural information processing systems*, 37:92529–92553, 2024. 2, 3

[53] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 3

[54] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 1

[55] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022. 6

[56] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European conference on computer vision*, pages 74–91. Springer, 2024. 3

[57] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 2

[58] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in neural information processing systems*, 36:13294–13307, 2023. 2, 3

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[60] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 1724–1732, 2024. 6