GuideSR: Rethinking Guidance for One-Step High-Fidelity Diffusion-Based Super-Resolution

Aditya Arora[1] ⋆ Zhengzhong Tu[2] † Yufei Wang[3] †
Ruizheng Bai[2] Jian Wang[3] ‡ Sizhuo Ma[3] ‡
[1] TU Darmstadt [2] Texas A&M University [3] Snap Inc.
aditya.arora@tu-darmstadt.de {tzz,rzbai}@tamu.edu {ywang25,jwang4,sma}@snapchat.com

Abstract

In this paper, we propose GuideSR, a novel single-step diffusion-based image super-resolution (SR) model specifically designed to enhance image fidelity. Existing diffusion-based SR approaches typically adapt pre-trained generative models to image restoration tasks by adding extra conditioning on a VAE-downsampled representation of the degraded input, which often compromises structural fidelity. GuideSR addresses this limitation by introducing a dual-branch architecture comprising: (1) a Guidance Branch that preserves high-fidelity structures from the original-resolution degraded input, and (2) a Diffusion Branch, which a pre-trained latent diffusion model to enhance perceptual quality. Unlike conventional conditioning mechanisms, our Guidance Branch features a tailored structure for image restoration tasks, combining Full Resolution Blocks (FRBs) with channel attention and an Image Guidance Network (IGN) with guided attention. By embedding detailed structural information directly into the restoration pipeline, GuideSR produces sharper and more visually consistent results. Extensive experiments on benchmark datasets demonstrate that GuideSR achieves state-of-the-art performance while maintaining the low computational cost of single-step approaches, with up to 1.39dB PSNR gain on challenging real-world datasets. Our approach consistently outperforms existing methods across various reference-based metrics including PSNR, SSIM, LPIPS, DISTS and FID, further representing a practical advancement for real-world image restoration.

1. Introduction

Image super-resolution (SR) aims to reconstruct high-resolution (HR) images from their low-resolution (LR) counterparts [4, 7, 20, 23, 24, 41, 51, 54, 55]. Beyond merely refining visible structures, SR seeks to recover details lost due to resolution degradation. This characteristic inherently renders SR an ill-posed problem, as multiple plausible HR images may correspond to the same LR input, necessitating models to accurately infer missing high-frequency details while maintaining fidelity to the original, often degraded content. Traditional SR methods have addressed this challenge using handcrafted priors [8, 9], regression-based deep neural networks [7, 17], as well as generative adversarial networks [19, 41].

In recent years, diffusion models have emerged as powerful frameworks for high-quality image generation [10, 30,

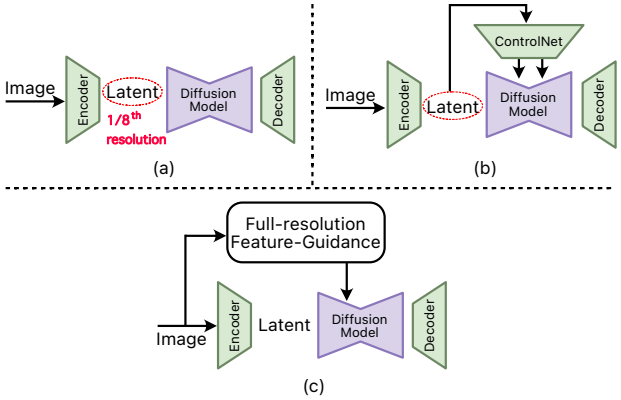


Figure 1. **Architecture comparison of diffusion-based super-resolution approaches**. (a) Standard diffusion process (e.g., OSEDiff [46]) processes latent representations directly; (b) Controller-based methods (e.g., DiffBIR [26], SeeSR [45], StableSR [40]) employ conditional mechanisms to guide the diffusion process; (c) Our proposed GuideSR introduces a dual-branch architecture with full-resolution feature guidance, preserving high-frequency details from the original input while leveraging the generative capabilities of diffusion models. This approach addresses the key limitation of existing methods the loss of structural fidelity due to VAE encoding of degraded inputs.

⋆ Work done during internship at Snap Inc.
† Equal Contribution ‡ Co-corresponding Author

32, 33]. These models operate through an iterative denoising process that gradually transforms random noise into coherent images, enabling high-quality image generation with unprecedented detail and diversity. It has been shown that such generative capabilities can also serve as a strong prior for recovering missing details in SR [34, 40] and other low-level vision tasks [15, 21, 26]. However, early diffusion-based methods typically require numerous denoising steps (e.g., 50-200) to recover a single image, significantly limiting their practicality. Recent advancements have proposed single-step restoration [43, 46], offering a promising direction for efficient and realistic super-resolution under real-world degradations.

While single-step diffusion SR methods [43, 46] greatly improve inference efficiency, they still struggle to preserve structural fidelity alongside realistic texture generation [39, 49]. We hypothesize that this limitation arises from the way existing methods condition their diffusion processes on LR inputs. As illustrated in Figure 1, current approaches typically encode the LR input to a latent space via a pretrained variational autoencoder (VAE). Then they either directly feed this LR latent into the denoising UNet (*e.g.* OSEDiff [46]), or condition on the LR latent through a controller mechanism [52] (*e.g.* DiffBIR [26], SeeSR [45], and StableSR [40]). Yet, since the VAE often employs aggressive downsampling with a high compression ratio (*e.g.* 8x [32]), it inevitably leads to the loss of high-frequency spatial details. Furthermore, because VAEs are primarily trained on high-quality images [46], their encoding process can degrade structural integrity when applied to lower-quality LR inputs. As a result, these approaches frequently exhibit suboptimal performance, particularly in reconstructing complex textures and fine patterns, failing to effectively balance detail hallucination and faithful reconstruction.

To address these limitations, we rethink the guidance mechanism design by taking an *SR-first* approach for diffusion-based SR modeling. We introduce GuideSR, a novel single-step diffusion-based SR method with a dual-branch architecture specifically designed to overcome the structural fidelity challenges on which most previous approaches struggle. As illustrated in Figure 1, our architecture consists of two complementary branches: **(1)** a Guidance Branch operating at full resolution to preserve structural details, and **(2)** a Diffusion Branch, which leverages the generative capabilities of a pre-trained latent diffusion model to enhance perceptual quality. Departing from traditional reliance on VAEs and controllers [32, 52], the Guidance Branch is specifically tailored to restoration, which directly processes the *full-resolution* low-quality input, bypassing the VAE latent encoding that causes detail loss in previous methods. This branch consists of Full Resolution Blocks (FRB) with Channel Attention mechanisms arranged in a residual-in-residual structure, along with an Image Guidance Network (IGN) that uses guided attention to preserve high-frequency details. Together, these components ensure effective feature extraction for high-fidelity reconstruction. Operating at the original image resolution allows this branch to preserve fine textures and structural information that would otherwise be lost in latent space processing. The main contributions of this work are:

- We introduce GuideSR, a novel single-step diffusion-based SR framework featuring a dual-branch architecture that effectively balances structural fidelity with perceptual quality, addressing a fundamental limitation in existing diffusion-based restoration methods.
- We design innovative Full Resolution Blocks and Image Guidance Network that adaptively refines full-resolution features, ensuring high-frequency details are preserved throughout the restoration process.
- We demonstrate, through extensive experiments, that our approach achieves state-of-the-art performance on multiple benchmarks, significantly surpassing existing methods in terms of fidelity and perceptual quality while maintaining computational efficiency.

2. Related Work

2.1. Efficient Diffusion Models

While Denoising Diffusion Probabilistic Models (DDPMs) [10] demonstrate exceptional image generation capabilities, they require 1000 training steps and typically 50 inference steps, which is prohibitive for many real-world applications. Latent Diffusion Models (LDM) and Stable Diffusion (SD) [32] enables generation of high-resolution images through a more computationally feasible latent space, yet they continue to depend on an iterative denoising process with billions of parameters. To mitigate this, faster samplers [6, 27, 38] have been developed to reduce the number of denoising steps. Recently, step distillation methods [29, 35] have emerged, distilling a pretrained diffusion model to significantly fewer steps. Notably, SD Turbo introduced Adversarial Diffusion Distillation (ADD) [37], facilitating single-step image synthesis. Beyond reducing the number of steps, techniques such as architecture optimization [11, 16] and quantization [13, 22] have been implemented to further improve the efficiency of diffusion models, making photorealistic image generation feasible on smartphones.

2.2. Diffusion Models for Image Super Resolution

Diffusion models have shown promising results in image super-resolution (SR). Early works like SR3 [34] adapted DDPM frameworks to SR tasks, demonstrating significant improvements in perceptual quality. StableSR [40] was the first to leverage prior knowledge in a pre-trained text-to-image diffusion model. DiffBIR [26] introduced a two-

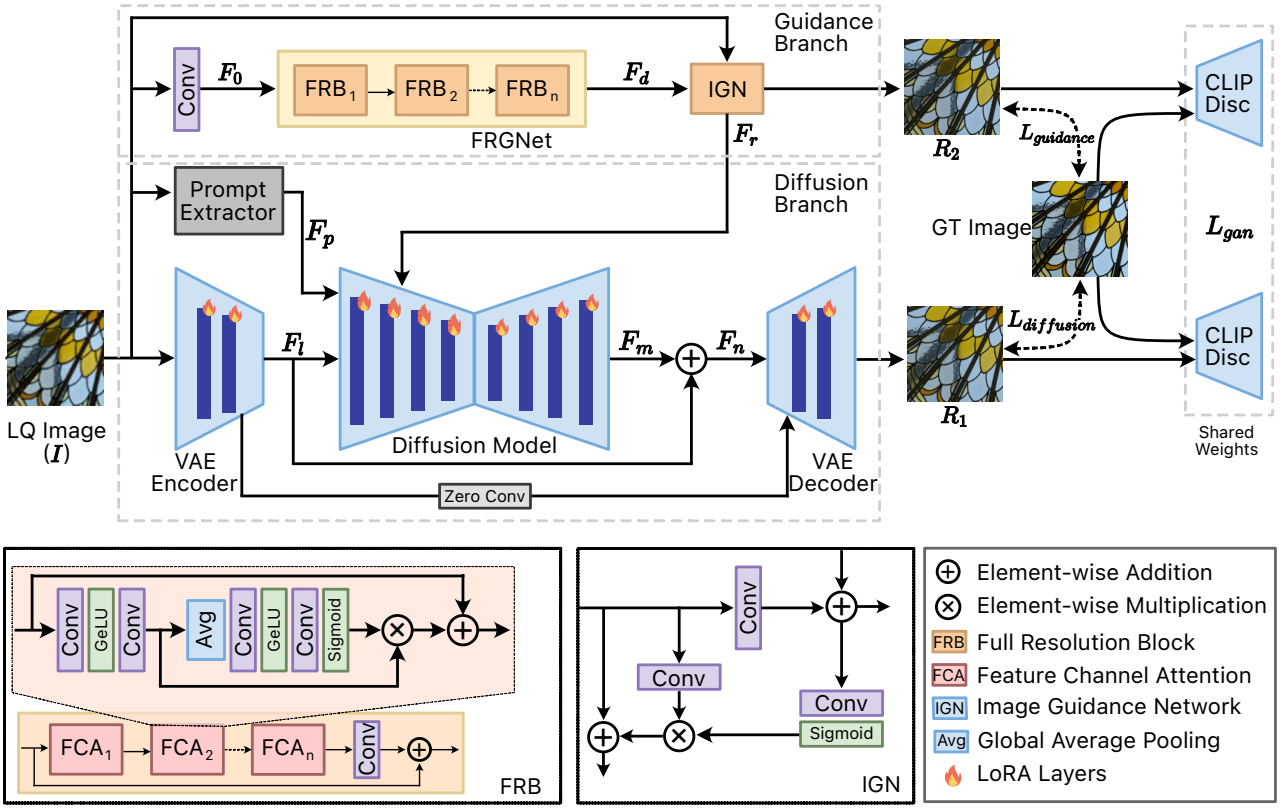


Figure 2. **Overview of GuideSR architecture.** Our method introduces a dual-branch architecture where the Guidance Branch (top) processes full-resolution input to preserve high-frequency details, while the Diffusion Branch (bottom) operates in the latent space for enhanced perceptual quality. The Guidance Branch applies a series of Full Resolution Blocks (FRBs) and an Image Guidance Network (IGN) to output a refined image R_2, which are enabled by the channel attention mechanisms and feature refinement operations detailed in the bottom. The Diffusion Branch employs a LoRA-finetuned diffusion model to produce the final output R_1, where the features from the Guidance Branch are adaptively refined and integrated into the denoise U-Net. Both branches are supervised through discriminators with shared weights during training.

stage model for degradation removal and realistic image reconstruction. SeeSR [45] utilized semantic prompts to generate detailed and semantically accurate results. CoDi [28] presented a conditional diffusion distillation method to accelerate diffusion sampling. PASD [49] proposed pixel-aware cross attention to effectively inject the pixel-level information into the diffusion model. ResShift [50] introduced a residual shift mechanism that significantly reduces the number of denoise steps. SinSR [43] further simplifies ResShift to a single-step model via consistency preserving distillation. OSEDiff [46] achieves single-step inference through LoRA-finetuning a pre-trained Stable Diffusion model with variational score distillation. Despite these advances, preserving structural information remains a challenge for diffusion-based SR models. Our work addresses this challenge by explicitly integrating structure information from the full-resolution LR input using a Guidance Branch specifically designed for image restoration, achieving higher fidelity compared to previous techniques.

3. Methodology

Our goal is to develop a single-step SR model that leverages the generative capabilities of diffusion priors without compromising the structural integrity of input images. To achieve this, we introduce GuideSR, a dual-branch architecture consisting of a Guidance Branch for efficient full-resolution feature- and image-level guidance and an enhanced Diffusion Branch for high-quality details synthesis. These branches work in tandem to refine degraded inputs while maintaining computational efficiency. In the following sections, we detail the architecture and operation of each branch (Sections 3.1 and 3.2), and describe our training strategy (Section 3.3).

3.1. Guidance Branch

Existing diffusion-based SR models rely on VAE and ControlNet to condition a diffusion model on a degraded input image, which effectively capture the image context but

struggle to retain fine-grained textures that are crucial for high-quality image restoration. To address this limitation, our Guidance Branch operates at the original image resolution, preserving structural and textural integrity.

Given a degraded image $I \in \mathbb{R}^{H\times W\times 3}$, we first apply a convolutional layer to extract low-level feature embeddings:

$$F_0 = \text{Conv}(I), \quad F_0 \in \mathbb{R}^{H\times W\times C}, \tag{1}$$

where $H \times W$ represents the spatial dimensions, and C denotes the number of feature channels. These shallow features F_0 are then passed through multiple Full Resolution Blocks (FRBs) to obtain deep features:

$$\begin{aligned} F_d &= \text{FRGNet}(F_0) \\ &= \text{FRB}_n \circ \ldots \circ \text{FRB}_2 \circ \text{FRB}_1(F_0), \quad F_d \in \mathbb{R}^{H\times W\times 2C}, \end{aligned} \tag{2}$$

where $\circ$ denotes function composition and FRB_i represents the i-th Full Resolution Block in the sequence.

As shown in Figure 2, each FRB consists of multiple Feature Channel Attention (FCA) blocks and skip connection arranged in a residual-in-residual structure. The detailed structure of an FRB, shown in the bottom-left of Figure 2, includes a standard channel attention pathway. Each FRB implements a residual connection where the input is added to a feature-recalibrated version of itself. The recalibration applies convolution, GeLU activation, average pooling, another convolution with sigmoid activation, and element-wise multiplication with the input, followed by a final convolution. This design enables effective feature representation by adaptively emphasizing crucial image regions while suppressing less relevant details.

To further refine the features, we employ an Image Guidance Network (IGN) which leverages guided attention to enhance structural integrity, as shown in Figure 2 (bottom-center). The IGN applies a residual connection where the input features F_d are added to an attention-modulated version of themselves. The attention mechanism involves applying convolutions and sigmoid activation to generate attention maps, which are then multiplied with another convolutional projection of the input features.

The IGN outputs two key components: a refined residual image R_2, which contributes to loss computation during training, and enriched feature representations F_r containing structural information. The refined residual image is obtained by adding a convolutional projection of the refined features to the original input image: $R_2 = \text{Conv}(F_r) + I$.

To align the spatial resolution of the refined features with the diffusion model's UNet-Encoder features, we apply pixel-unshuffle operations for downsampling:

$$F_r' = \text{PixelUnshuffle}(F_r, s) \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times s^2\cdot 2C}, \tag{3}$$

where s is the downsampling factor. The pixel-unshuffle operation rearranges the spatial dimensions of a tensor while preserving all values, making it ideal for the SR task.

The downsampled features F_r' are then concatenated with UNet encoder outputs to enrich hierarchical feature representations at multiple scales, as illustrated by the vertical connections in Figure 2. This multi-scale integration allows structural information to influence the diffusion process at various resolutions, preserving details that might otherwise be lost.

3.2. Diffusion Branch

We use a pretrained latent diffusion model as our generative prior, as shown in Figure 2 (bottom branch). Starting from the high-resolution input, the VAE Encoder progressively reduces spatial dimensions while expanding channel capacity. The encoder transforms the input image I into latent features $F_l = \text{VAE}_{\text{Encoder}}(I) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 4C}$.

Additionally, a text prompt is extracted from the input image using a dedicated prompt extraction module. This produces a text embedding vector $F_p = \text{PromptExtractor}(I)$, allowing the diffusion model to leverage text-conditional generation capabilities and guide the restoration process with semantic understanding.

The central UNet structure processes these latent features with the integrated guidance from both the text prompt and the downsampled Guidance Branch features. While the pretrained diffusion model is conditioned on the timestep t, we choose a fixed t_f for the one-step model:

$$F_m = \text{UNet}(F_l, t_f, F_p, \{F_r'\}) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 4C}, \tag{4}$$

where $\{F_r'\}$ represents the set of multi-scale features from the Guidance Branch. Notice that a long-skip connection is added such that the UNet only predicts the *residual* latent, with the final latent features computed as:

$$F_n = F_m + F_l. \tag{5}$$

This skip connection ensures that low-frequency information from the original input is preserved throughout the diffusion process, allowing the network to focus on generating high-frequency details while maintaining overall image structure. The output of the UNet is then decoded through the VAE Decoder to produce the result $R_1 = \text{VAEDecoder}(F_n) \in \mathbb{R}^{H\times W\times 3}$.

During training, LoRA layers [12] are added for parameter-efficient finetuning. For a weight matrix $W \in \mathbb{R}^{d\times k}$ in the original network, LoRA parameterizes the update as a low-rank decomposition $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$ are the ranks. Specifically, we employ LoRA ranks of $r = 8$ for UNet layers and $r = 4$ for the VAE. We also add skip-connections with Zero-Convs between the VAE Encoder and Decoder [31] to mitigate the detail loss due to VAE in the Diffusion Branch.

Table 1. **Quantitative comparison of diffusion-based super-resolution methods** across DIV2K-Val, DRealSR, and RealSR datasets. GuideSR consistently outperforms both multi-step methods (requiring 15-200 steps) and recent one-step approaches across all reference-based metrics (PSNR/SSIM/LPIPS/DISTS/FID). Our method achieves significant improvements on the challenging real-world DRealSR dataset with a **1.39dB PSNR gain** over the best multi-step method (ResShift) while maintaining the efficiency of single-step inference. Bold values indicate the best results for each metric.

Dataset	Method	Steps	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	CLIPIQA ↑
DIV2K	StableSR [40]	200	23.26	0.5726	0.3113	0.2048	24.44	4.7581	65.92	0.6192	0.6771
	DiffBIR [26]	50	23.64	0.5647	0.3524	0.2128	30.72	4.7042	65.81	0.6210	0.6704
	SeeSR [45]	50	23.68	0.6043	0.3194	0.1968	25.90	4.8102	68.67	0.6240	**0.6936**
	PASD [49]	20	23.14	0.5505	0.3571	0.2207	29.20	**4.3617**	**68.95**	**0.6483**	0.6788
	ResShift [50]	15	24.65	0.6181	0.3349	0.2213	36.11	6.8212	61.09	0.5454	0.6071
	SinSR [43]	1	24.41	0.6018	0.3240	0.2066	35.57	6.0159	62.82	0.5386	0.6471
	OSEDiff [46]	1	23.72	0.6108	0.2941	0.1976	26.32	4.7097	67.97	0.6148	0.6683
	GuideSR	**1**	**24.76**	**0.6333**	**0.2653**	**0.1879**	**21.04**	5.9273	63.97	0.5679	0.5840
DRealSR	StableSR [40]	200	28.03	0.7536	0.3284	0.2269	148.98	6.5239	58.51	0.5601	0.6356
	DiffBIR [26]	50	26.71	0.6571	0.4557	0.2748	166.79	6.3124	61.07	0.5930	0.6395
	SeeSR [45]	50	28.17	0.7691	0.3189	0.2315	147.39	6.3967	**64.93**	0.6042	0.6804
	PASD [49]	20	27.36	0.7073	0.3760	0.2531	156.13	**5.5474**	64.87	**0.6169**	0.6808
	ResShift [50]	15	28.46	0.7673	0.4006	0.2656	172.26	8.1249	50.60	0.4586	0.5342
	SinSR [43]	1	28.36	0.7515	0.3665	0.2485	170.57	6.9907	55.33	0.4884	0.6383
	OSEDiff [46]	1	27.92	0.7835	0.2968	0.2165	135.30	6.4902	64.65	0.5899	**0.6963**
	GuideSR	**1**	**29.85**	**0.8078**	**0.2640**	**0.1960**	**122.06**	7.7500	57.14	0.5230	0.5762
RealSR	StableSR [40]	200	24.70	0.7085	0.3018	0.2288	128.51	5.9122	65.78	0.6221	0.6178
	DiffBIR [26]	50	24.75	0.6567	0.3636	0.2312	128.99	5.5346	64.98	0.6246	0.6463
	SeeSR [45]	50	25.18	0.7216	0.3009	0.2223	125.55	5.4081	**69.77**	0.6442	0.6612
	PASD [49]	20	25.21	0.6798	0.3380	0.2260	124.29	**5.4137**	68.75	**0.6487**	0.6620
	ResShift [50]	15	26.31	0.7421	0.3460	0.2498	141.71	7.2635	58.43	0.5285	0.5444
	SinSR [43]	1	26.28	0.7347	0.3188	0.2353	135.93	6.2872	60.80	0.5385	0.6122
	OSEDiff [46]	1	25.15	0.7341	0.2921	0.2128	123.49	5.6476	69.09	0.6326	**0.6693**
	GuideSR	**1**	**27.08**	**0.7681**	**0.2407**	**0.1878**	**96.83**	6.7647	62.20	0.5716	0.5482

3.3. Training Strategy

Previous work has shown that adversarial training help reduce the number of steps of diffusion models significantly [25, 37, 47, 48]. We utilizes dual discriminators (shown in blue in Figure 2) that evaluate both the Guidance Branch output R_2 and the Diffusion Branch output R_1. These discriminators share weights to ensure consistency in the adversarial training signal. With this adversarial training, we are able to leverage the pretrained diffusion prior and train our one-step restoration model by computing the following loss directly from the restored image and the ground truth:

$$\mathcal{L}_B = \lambda_1 \cdot \mathcal{L}_{\text{MSE}} + \lambda_2 \cdot \mathcal{L}_{\text{LPIPS}} + \lambda_3 \cdot \mathcal{L}_{\text{GAN}}, \tag{6}$$

where the individual loss terms are defined as:

$$\begin{aligned} \mathcal{L}_{\text{MSE}}(R, Y) &= \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} ||R_{i,j} - Y_{i,j}||_2^2 \\ \mathcal{L}_{\text{LPIPS}}(R, Y) &= ||\Phi(R) - \Phi(Y)||_2^2 \\ \mathcal{L}_{\text{GAN}}(R) &= -\mathbb{E}[\log(D(R))] \end{aligned} \tag{7}$$

where R is the restored image, Y is the ground truth, Φ represents the LPIPS network, and D is the discriminator network. The weighting coefficients are set as $\lambda_1 = 1.0$, $\lambda_2 = 5.0$, and $\lambda_3 = 0.5$.

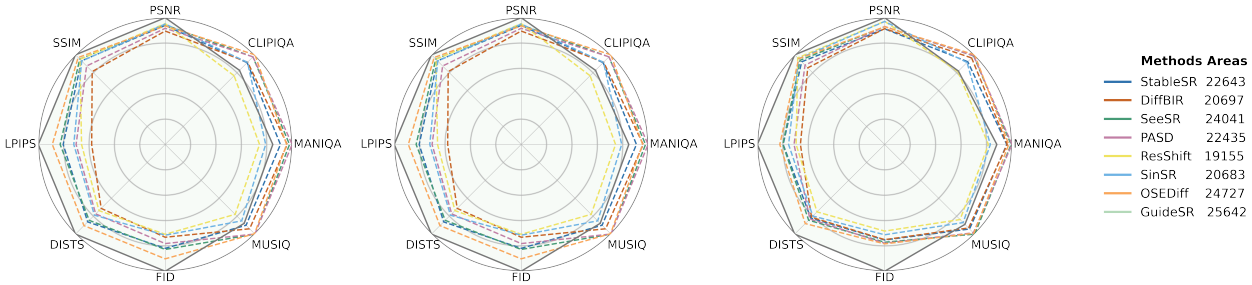
While we use the Diffusion Branch as the final output during inference, computing the loss on the Guidance Branch helps stabilize training. The final loss combines both branches with different weights:

$$\mathcal{L}_{\text{final}} = \lambda_d \cdot \mathcal{L}_B(Y, R_1) + \lambda_g \cdot \mathcal{L}_B(Y, R_2), \tag{8}$$

where Y is the ground truth image, R_1 is the output from the Diffusion Branch, and R_2 is the output from the Guidance Branch, with $\lambda_d = 0.9$ and $\lambda_g = 0.1$.

4. Results

Datasets. Follow the training setup in ResShift [50], we construct our training dataset using high-resolution (HR) images randomly cropped from ImageNet [5] and FFHQ [14]. The corresponding low-resolution (LR) images are synthesized using the degradation pipeline from



(a) RealSR (b) DRealSR (c) DIV2K

Figure 3. **Multi-metric performance visualization.** We visualize the performance of different SR methods, displaying reference-based metrics (PSNR, SSIM, LPIPS, DISTS, FID) on the left and no-reference metrics (CLIPIQA, MANIQA, MUSIQ) on the right. Our GuideSR model prioritizes fidelity, consistently achieving the best reference-based metrics across all datasets. It does not lead in no-reference metrics due to the perception-distortion tradeoff [2]. Despite this, GuideSR consistently covers the largest area across all datasets, demonstrating a superior balance across various quality aspects.

Real-ESRGAN [42], augmented with random horizontal flipping. The final HR target images are cropped to 256×256. For evaluation, we utilize the test sets provided by StableSR [40], which encompass both synthetic 3,000 512×512 images from DIV2K-val [1] with Real-ESRGAN degradation) and real-world images (paried LQ and 512×512 HQ images from RealSR [3] and DRealSR [44]).

Training Details. We use a pretrained Stable Diffusion Turbo (v2.1) [36] as our diffusion prior. RAM [56] is used as the prompt extractor. CLIP [18] is used as the discriminator. We train our model using the AdamW optimizer with an initial learning rate of 5×10^{-5}, following a cosine annealing schedule. To stabilize training, a warm-up phase of 500 iterations is applied, with a total iteration of 100k. All experiments are conducted on an 8-GPU cluster equipped with NVIDIA A100 GPUs, each with 40 GB of memory.

4.1. Quantitative Comparison

We evaluate GuideSR against several state-of-the-art diffusion-based SR methods on three benchmark datasets: DIV2K-Val, DRealSR, and RealSR. As shown in Table 1, our method consistently outperforms both multi-step (StableSR [40], DiffBIR [26], SeeSR [45], PASD [49], and ResShift [50]) and single-step (SinSR [43], OSEDiff [46]) approaches across all reference-based metrics including PSNR, SSIM, LPIPS, DISTS and the distribution metric FID.

Quantitative Results on DIV2K-Val. On the synthetic DIV2K-Val dataset, GuideSR achieves a PSNR of 24.76dB and an SSIM of 0.6333, surpassing the best previous method (ResShift) by 0.11dB and 0.0152 while requiring only a single inference step instead of 15. In terms of perceptual quality, our method demonstrates remarkable improvements with the lowest LPIPS (0.2653), DISTS (0.1879), and FID (21.04) scores among all methods.

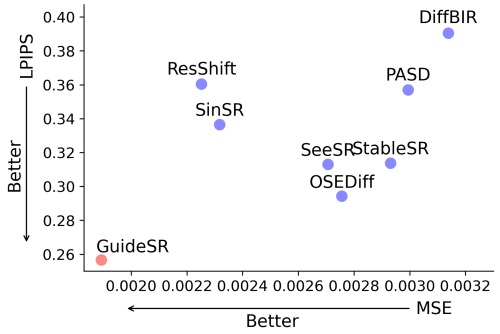


Figure 4. **Pixel-Space Fidelity (MSE) and Feature-Space Fidelity (LPIPS).** Enhancing both MSE and LPIPS is generally challenging because boosting generative capabilities often increases feature-space fidelity and enhances the realism of restored images, but usually at the cost of reduced pixel-space fidelity. GuideSR achieves the highest scores in both MSE and LPIPS among all methods, demonstrating its ability to maintain fidelity in both pixel and feature spaces.

Quantitative Results on Real-World Datasets. The advantages of GuideSR are even more pronounced on real-world datasets, which present more challenging and diverse degradations than synthetic ones. For example, on the DRealSR dataset, our method achieves a remarkable PSNR of 29.85dB, surpassing the best previous method (ResShift) by 1.39dB. The perceptual metrics also show significant improvements, with our method reducing the FID score to 122.06 on DRealSR (13.24 lower than the previous best) and 96.83 on RealSR (26.66 lower than the previous best). These consistent improvements across different real-world datasets demonstrate the robustness and generalizability of our approach.

Pixel-Space Fidelity and Feature-Space Fidelity. Figure 4 visualizes the MSE and LPIPS of different methods

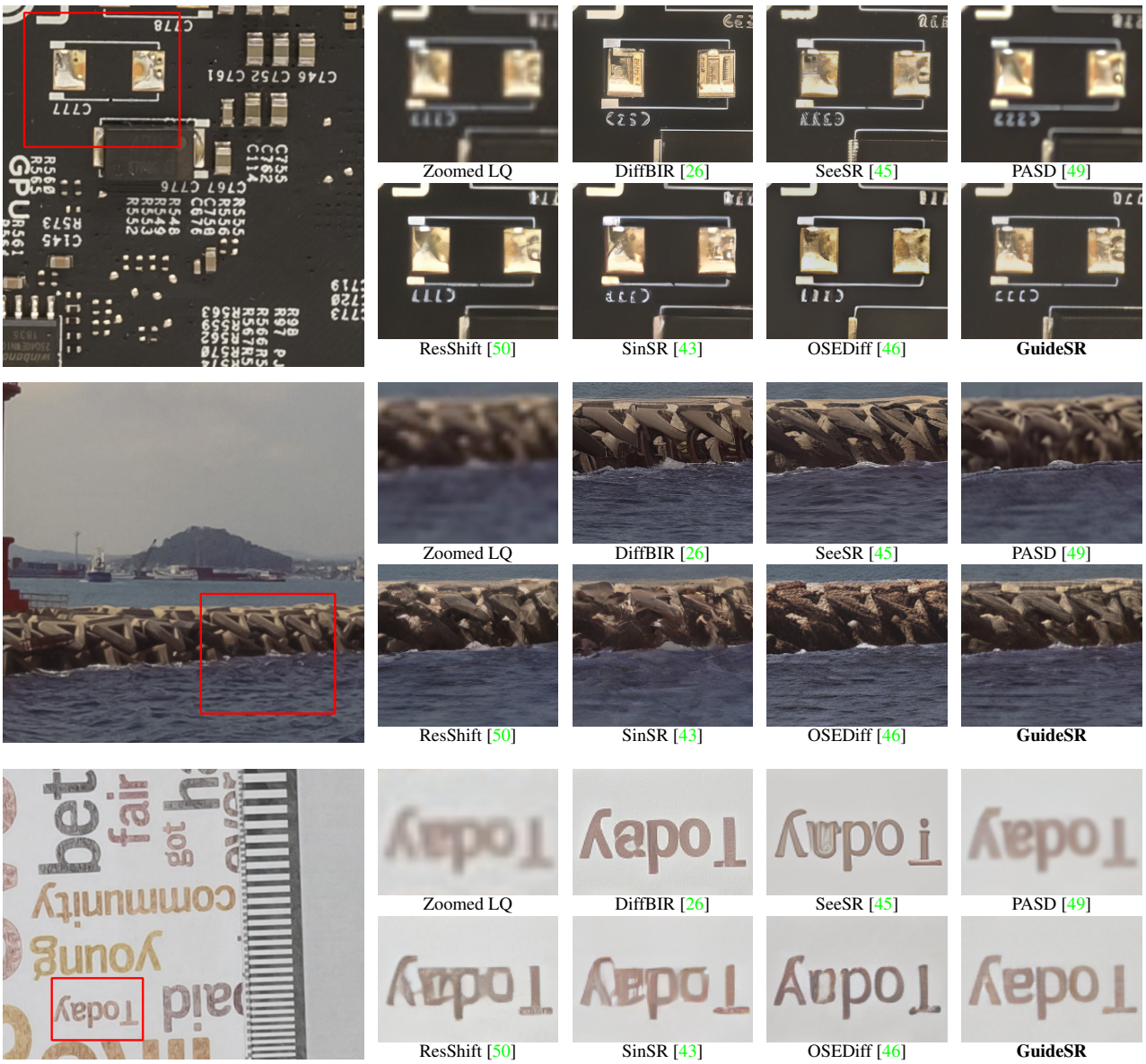


Figure 5. **Visual Comparison on Real-World Images from RealSR [3] and DRealSR [44].** (Top) GuideSR accurately restores detailed features such as the text shape and the reflections on metal surfaces. (Middle) GuideSR correctly reconstructs the geometry of the concrete blocks, while OSEDiff introduces incorrect textures and a color shift. (Bottom) GuideSR faithfully restores the text, particularly the inverted "a" letter, whereas PASD generates authentic but slightly blurred text.

on all three test sets. While MSE measures the fidelity in the pixel space, LPIPS measures the fidelity in the feature space and aligns well with human perception (perceptual similarity [53]). Achieving improvements in both MSE and LPIPS is typically challenging, as enhancing generative capabilities often leads to a decline in pixel-space fidelity while increasing the feature-space fidelity [50]. Nevertheless, GuideSR achieves the best scores in both MSE and LPIPS among all methods, demonstrating its ability to preserve both pixel-space and feature-space fidelity.

Full-Reference and No-Reference Metrics. Since GuideSR focuses on the fidelity of the restored image, its performance is best evaluated by the full-reference metrics discussed above. That said, we also report no-reference image quality assessment (IQA) metrics including NIQE, MUSIQ, MANIQA and CLIPIQA for completeness. GuideSR does not achieve the best performance on these metrics because of the perception-distortion tradeoff:

Blau and Michaeli [2] proved mathematically that less distortion (better full-reference scores[1]) and better perceptual quality (better no-reference scores) cannot be achieved by the same image restoration algorithm. Given that GuideSR achieves lowest distortion on all the three datasets across all full-reference metrics, it is unsurprising that it scores lower on no-reference metrics. Nevertheless, it performs comparably to existing methods.

Multi-Metric Performance Visualization. To jointly compare the full-reference and no-reference metrics, Figure 3 visualizes the performance of different methods across all metrics using spider charts for each dataset. These radar plots provide an intuitive comparison where larger area coverage indicates better overall performance across metrics. As shown in the figure, GuideSR consistently covers the largest area across all three datasets, achieving superior balance across different quality aspects. See the supplementary material for the details of this visualization.

4.2. Qualitative Results

Figure 5 presents visual comparisons between GuideSR and other state-of-the-art methods on the RealSR dataset. The top row shows GuideSR's ability to reconstruct electronic components with intricate patterns, capturing small details more accurately than other approaches, which often miss fine details or introduce false patterns. The middle row illustrates the reconstruction of triangular concrete blocks near a sea. While other methods struggle to accurately capture the exact block structure or unintentionally generate false textures, our method achieves most precise restoration of the complex geometry. OSEDiff also gives close geometry but introduces incorrect textures and a color shift. The bottom row depicts the reconstruction of text. Although the proposed model is not specifically trained on text images, the guidance branch enables it to restore text with the greatest accuracy, especially for the inverted "a" letter. PASD also generates the correct letters but the result looks slightly blurry. These qualitative results validate the effectiveness of our approach in preserving high-frequency details and structural integrity while enhancing perceptual quality, making GuideSR particularly suitable for real-world applications where both fidelity and visual quality are essential. **See the supplementary material for more qualitative results.**

4.3. Ablation Study

To evaluate the contribution of each component, we conducted an ablation study using RealSR [3]. The results are presented in Table 2. "Baseline" refers to using the Diffusion Branch only without the long-skip connections.

[1]Notice that although feature-space metrics like LPIPS and DISTS highly correlate to human perception, they are full-reference metrics and are not considered perception metrics by definition [2].

Experiment	PSNR ↑
Baseline	26.65
Baseline + Long-skip	26.80
Baseline + Guidance	26.82
Baseline + Guidance + IGN + Long-skip	27.08

Table 2. Ablation study on RealSR [3] showing the contribution of different components to GuideSR's performance.

Effect of Long-Skip Connections. Adding long-skip connections to the baseline model improves the PSNR from 26.65dB to 26.80dB, demonstrating the importance of preserving low-frequency information throughout the network.

Effect of Guidance Branch. Incorporating the Guidance Branch provides a PSNR improvement from 26.65dB to 26.82dB over the baseline. This confirms our hypothesis that explicit structural guidance from the full-resolution input significantly enhances reconstruction quality.

Effect of Image Guidance Network (IGN). The complete architecture (Baseline + Guidance + IGN + Long-skip) achieves the best performance with a PSNR of 27.08dB, which is 0.43dB higher than the baseline. This substantial improvement highlights the effectiveness of IGN in adaptively refining features and ensuring high-frequency details are preserved throughout the restoration process.

5. Conclusion

In this study, we propose a super-resolution-centric framework designed to address the fidelity challenges present in existing one-step diffusion-based SR models. Our dual-branch architecture includes a Diffusion Branch with strong generative capabilities and a Guidance Branch that effectively extracts detailed structural and textural features. By utilizing the full-resolution degraded input as a guide, the Guidance Branch significantly enhances the fidelity of the reconstruction process while maintaining the computational efficiency of one-step SR models.

Limitations and future works. Although one-step approaches significantly reduce inference time, the size of the denoising UNet still limits the checkpoint size and memory footprint, posing challenges for resource-constrained devices such as smartphones. Future work might focus on further improving efficiency through architectural optimization [11, 16] and quantization [13, 22]. Another potential direction is to extend the proposed framework to video restoration, where maintaining fidelity and computational efficiency is also crucial.

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.

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

[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.

[4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems*, 35:30150–30166, 2022.

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.

[8] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (ToG)*, 30(2):1–11, 2011.

[9] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009.

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

[11] Dongting Hu, Jierun Chen, Xijie Huang, Huseyin Coskun, Arpit Sahni, Aarush Gupta, Anujraaj Goyal, Dishani Lahiri, Rajesh Singh, Yerlan Idelbayev, et al. Snapgen: Taming high-resolution text-to-image models for mobile devices with efficient architectures and training. *arXiv preprint arXiv:2412.09619*, 2024.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[13] Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. Tfmq-dm: Temporal feature maintenance quantization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7362–7371, 2024.

[14] 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.

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[16] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion. In *European Conference on Computer Vision*, pages 381–399. Springer, 2024.

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[18] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10651–10662, 2022.

[19] 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.

[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.

[21] Jinlong Li, Baolu Li, Zhengzhong Tu, Xinyu Liu, Qing Guo, Felix Juefei-Xu, Runsheng Xu, and Hongkai Yu. Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15205–15215, 2024.

[22] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545, 2023.

[23] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[24] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022.

[25] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

[26] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[28] Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M Patel, and Peyman Milanfar. Codi: conditional diffusion distillation for higher-fidelity and faster image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9048–9058, 2024.

[29] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[30] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[31] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.

[32] 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.

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[34] 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.

[35] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[36] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[37] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024.

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[39] Yuhao Wan, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, Ming-Ming Cheng, and Bo Li. Clearsr: Latent low-resolution image embeddings help diffusion-based real-world super resolution models see clearer. *arXiv preprint arXiv:2410.14279*, 2024.

[40] 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*, pages 1–21, 2024.

[41] 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 (ECCV) workshops*, pages 0–0, 2018.

[42] 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.

[43] 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. *arXiv preprint arXiv:2311.14760*, 2023.

[44] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020.

[45] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. *arXiv preprint arXiv:2311.16518*, 2023.

[46] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024.

[47] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

[48] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8196–8206, 2024.

[49] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023.

[50] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *arXiv preprint arXiv:2307.12348*, 2023.

[51] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021.

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[53] 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.

[54] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, pages 649–667. Springer, 2022.

[55] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[56] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.