

Adversarial Score Distillation for Stable One-Step Diffusion in Real-World Image Super-Resolution

Wei Zhu¹, Kai Zhang^{2*}, Yu Zheng¹, Lei Luo^{1*}, and Jian Yang^{1,2*}

¹ PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology

² PCA Lab, School of Intelligence Science and Technology, Nanjing University

Abstract. In recent years, super-resolution (SR) methods leveraging pre-trained text-to-image diffusion models have shown impressive potential. However, their deployment in real-world applications remains constrained by the computational overhead of numerous inference steps. Recent one-step approaches, such as OSediff and TSD-SR, mitigate this issue via score distillation techniques. However, their discriminators are optimized solely with the diffusion loss, resulting in weak and unstable gradients for the generator, making it difficult to preserve fine details and structural fidelity. To tackle these limitations, we introduce ASDSR, a novel one-step distillation framework tailored for real-world image super-resolution. Our approach enhances Adversarial Score Distillation (ASD) by adopting a WGAN-based formulation with a stabilized penalty term that utilizes the full discriminator loss to improve the distillation process. Furthermore, we incorporate CLIP-derived semantic priors to align reconstructed images with their high-quality counterparts in the feature space, effectively guiding the generator toward perceptual realism. Extensive experiments on multiple real-world benchmarks demonstrate that ASDSR achieves superior performance compared to state-of-the-art methods, offering a practical and efficient solution for real-world image restoration.

Keywords: Super-Resolution · Diffusion Models · Score Distillation

1 Introduction

Image super-resolution (ISR) [5, 23, 39, 61] aims to reconstruct a high-quality (HQ) image from its low-quality (LQ) image counterpart that suffers from various degradations. Different from traditional non-blind ISR [18, 20], which assumes an ideal bicubic downsampling kernel, real-world image super-resolution (Real-ISR) [10, 38, 58, 59] is designed to generate HQ images from corresponding LQ images affected by various unknown and complex degradations.

Generative models, particularly Generative Adversarial Networks (GANs) [9, 37, 54] and Diffusion Models (DMs) [13, 33], have demonstrated remarkable

* Corresponding authors (kaizhang@nju.edu.cn, cslluo, csjyang@njust.edu.cn).

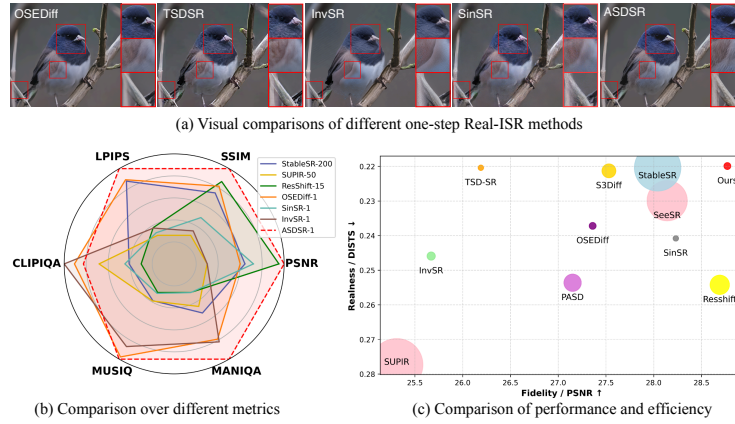


Fig. 1: (a) Visual comparison between our ASDSR and other state-of-the-art Real-ISR methods. (b) Performance comparison on the DrealSR benchmark. All evaluation scores are normalized for visualization. (c) Model efficiency comparison. Larger bubbles indicate longer inference time, and bubbles closer to the top-right represent better fidelity and realism.

performance in handling Real-ISR tasks. GAN-based methods leverage adversarial training to produce highly realistic images, but the inherent instability of this process often results in unnatural artifacts. Recently, diffusion-based ISR models [24, 35, 53] leverage the strong image priors from pre-trained text-to-image (T2I) models, leading to more realistic detail generation compared to GAN-based approaches. However, the iterative denoising of diffusion models makes Real-ISR computationally expensive and prone to unwanted randomness.

To achieve an efficient one-step network similar to GANs, several pioneering methods have focused on distilling diffusion models to reduce their iterative steps [6, 40, 46, 48]. SinSR [40] distills the 15-step ResShift [56] into a one-step ISR model. However, due to the lack of large pre-trained T2I models, it tends to produce over-smoothed results with limited perceptual realism. OSEDiff [46] leverages Variational Score Distillation (VSD) [41] to reduce the number of inference steps by using the LQ image as the starting point, and achieves state-of-the-art performance on Real-ISR tasks. Nevertheless, the unreliable gradient directions introduced by the VSD loss limit its ability to recover fine image details [6]. To compensate for the limitation of the VSD loss, TSD-SR [6] further extends this paradigm with Target Score Distillation (TSD) to improve perceptual realism, but the additional Low-Rank Adaptation (LoRA) [15] branch inevitably increases training complexity and slows convergence. Beyond these limitations, both OSEDiff and TSD-SR still follow the VSD paradigm, where the discriminator is updated solely through diffusion loss, yielding weak and unstable gradients that fail to provide effective guidance for the generator, thus limiting fine-detail reconstruction.

To address the above challenges, we propose ASDSR, a novel one-step diffusion model tailored to the Real-ISR task. Specifically, ASDSR adopts two key

strategies to ensure realistic and natural reconstruction. First, we extend Adversarial Score Distillation (ASD) [44] under the WGAN [1] paradigm by employing the complete discriminator loss with a stabilized penalty term to enhance training stability and detail fidelity. Second, we improve perceptual naturalness by integrating CLIP guidance to enforce semantic consistency between the reconstructed HQ images and their corresponding ground truth. As illustrated in Fig. 1, ASDSR demonstrates superior performance in both reconstruction quality and inference efficiency. The main contributions of this work are summarized as follows:

- We redesign Adversarial Score Distillation with a stabilized formulation tailored for one-step diffusion-based super-resolution, significantly improving training robustness and perceptual realism.
- To enforce semantic consistency and enhance perceptual naturalness, we leverage the visual semantic priors of CLIP to align the reconstructed images with real-world HQ images in the CLIP semantic space.
- Extensive experiments on both synthetic and real-world benchmarks demonstrate that ASDSR achieves superior restoration performance and high efficiency.

2 Related Work

Score Distillation. Score distillation was introduced by Score Distillation Sampling (SDS) [30] as a method to optimize representations by distilling the generative priors of pre-trained text-to-image diffusion models. Although effective, SDS often suffers from over-smoothing and limited diversity due to its reliance on fixed diffusion gradients. To address these issues, ProlificDreamer [41] proposed Variational Score Distillation (VSD), which employs Wasserstein gradient flow to stabilize optimization and achieve more diverse results. Recent works have extended score distillation to diffusion distillation [8, 27]. Distribution Matching Distillation [51, 52] refines and extends VSD by minimizing an approximate KL divergence. Its gradients are computed from the differences between fixed and dynamically trained score functions, thus enabling the efficient training of high-fidelity one-step generative models. Furthermore, Adversarial Score Distillation (ASD) [44] elucidates the relationship between score distillation and WGAN [1], and exploits powerful diffusion models in the form of score distillation. Building on these advances, we reinterpret ASD from a WGAN perspective and introduce a stabilized penalty formulation that enables efficient one-step training for Real-ISR.

One-step Diffusion-based Real-ISR. Recently, Stable Diffusion (SD) models pre-trained on T2I tasks have demonstrated remarkable performance across various real-world visual applications. Due to their strong natural image priors and versatility, these models have also been extensively adopted for Real-ISR tasks [24, 50, 53]. However, the multi-step denoising process increases computational cost and raises the risk of synthesizing unfaithful content. To address this

limitation, several methods have been proposed to distill the multi-step diffusion process into a more efficient one-step student model [6, 41, 44, 46, 52]. For instance, OSediff [46] directly uses the LQ image as the starting point of the diffusion process and employs the VSD loss as a regularizer to accelerate inference. S3Diff [57] designs a degradation-guided LoRA module and an online negative prompting strategy to enhance restoration quality. TSD-SR [6] leverages Target Score Distillation to produce more reliable gradients and improve the realism of generated outputs. However, all the aforementioned methods still fall short of achieving semantic consistency and preserving faithful details.

3 Methodology

3.1 Preliminaries.

Problem Formulation. The objective of Real-ISR is to reconstruct an HQ image x_H from a given LQ input x_L . With recent advances in deep learning, training a parameterized neural network G_θ from a training dataset S of (x_L, x_H) pairs has become a common approach to mapping LQ images to their HQ counterparts. Mathematically, this can be formulated as minimizing the following objective:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x_L, x_H) \sim S} [\mathcal{L}_{rec}(G_\theta(x_L), x_H) + \lambda \mathcal{L}_{reg}(q_\theta(G_\theta(x_L)), p(x_H))], \quad (1)$$

where $\mathcal{L}_{rec}$ and $\mathcal{L}_{reg}$ denote reconstruction and regularization losses, respectively. $\mathcal{L}_{rec}$ is typically measured using distance metrics such as L_1 norm, L_2 norm and LPIPS [60]. The regularization term $\mathcal{L}_{reg}$ aims to improve the realism and naturalness of the generated HQ image, denoted by $\hat{x}_H$, and λ is a balance parameter. Here, $p(x_H)$ represents the distribution of real-world HQ images, while $q_\theta(\hat{x}_H)$ denotes the distribution of restored HQ images. The objective of $\mathcal{L}_{reg}$ can thus be viewed as aligning $p(x_H)$ and $q_\theta(\hat{x}_H)$ by minimizing the Kullback-Leibler (KL) divergence [19]:

$$\min_{\theta} \mathcal{D}_{KL}(q_\theta(\hat{x}_H) || p(x_H)). \quad (2)$$

Existing studies [38, 58] optimize this objective via adversarial training. However, because the discriminator is trained from scratch alongside the generator, it may fail to capture the full HQ image distribution, leading to insufficient discriminative capability and issues such as mode collapse and training instability. Recent T2I diffusion models [29, 33], trained on billions of image-text pairs, can effectively capture the distribution of natural images. Building on this, recent works [6, 46] employ Variational Score Distillation (VSD) [41, 52] to align the distributions of restored and natural images, achieving competitive performance.

Score Distillation. VSD employs KL divergence as the objective to align the distributions of two diffusion models. A pre-trained diffusion model, represented as ϵ_ϕ , and its trainable LoRA model ϵ_ψ are used to regularize the generator G_θ .

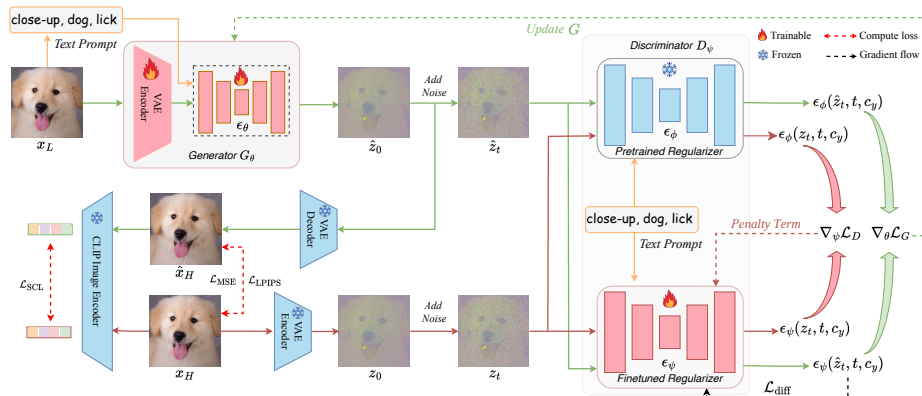


Fig. 2: The training framework of ASDSR. We train a one-step generator G_θ to map the LQ image x_L to the desired HQ image. The noisy latent $\hat{z}_t$ and z_t will be fed into the discriminator D_ψ , which consists of a frozen pre-trained regularizer ϵ_ϕ and a trainable regularizer ϵ_ψ . Adversarial score distillation is then performed to provide gradient guidance $\mathcal{L}_G$ for updating the generator G_θ . The regularizer ϵ_ψ is optimized separately using the diffusion loss $\mathcal{L}_{\text{diff}}$ and penalty term to stabilize adversarial learning.

The gradient with respect to the generator parameters θ is formulated as:

$$\nabla_{\theta} \mathcal{L}_{VSD}(\hat{z}, c_y) = \mathbb{E}_{t, \epsilon} [\omega(t) (\epsilon_{\psi}(\hat{z}_t, t, c_y) - [\epsilon_{\phi}(\hat{z}_t, t) + \lambda (\epsilon_{\phi}(\hat{z}_t, t, c_y) - \epsilon_{\phi}(\hat{z}_t, t)])] \frac{\partial \hat{z}}{\partial \theta}], \quad (3)$$

where $\hat{z}_t = \alpha_t \hat{z} + \sigma_t \epsilon$ is the noise latent input, $\hat{z}$ is the latent output by the generator network G_θ , $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise, and α_t, σ_t are scalars dependent on the diffusion timestep. c_y is a text embedding corresponding to a caption that describes the input image. $\omega(t)$ is a weight based on the time t , and λ is the classifier-free guidance (CFG) scale. The learnable LoRA regularizer ϵ_ψ is typically optimized using a standard diffusion objective. ASD [44] enhances this process with a full WGAN discriminator loss, achieving strong performance in image editing tasks. The update of ϵ_ψ is driven by the following objective:

$$\nabla_{\psi} \mathcal{L}_D(\hat{z}_t, c_y) = \nabla_{\psi} \mathbb{E}_{t, \epsilon} [\|\epsilon_{\psi}(\hat{z}_t, t, c_y) - \epsilon\|^2 + \gamma \|\epsilon_{\phi}(\hat{z}_t, t, c_y) - \epsilon_{\psi}(\hat{z}_t, t, c_y)\|^2], \quad (4)$$

where γ is an adjustable hyperparameter that controls the strength and sign of the regularization term.

3.2 Overview of Our Method

Framework Overview. Our objective is to design a one-step diffusion network tailored for Real-ISR, ensuring that restored HR images achieve maximum naturalness while maintaining consistency with the LR inputs. As depicted in Fig. 2, our pipeline consists of two key components: a one-step generator G_θ and a trainable discriminator D_ψ . The generator G_θ comprises a trainable encoder E_θ , a fine-tuned diffusion network ϵ_θ , and a frozen decoder D . The discriminator D_ψ

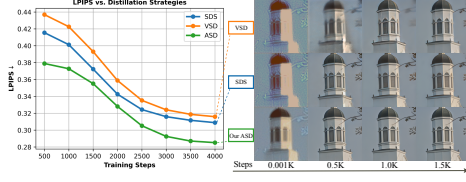


Fig. 3: LPIPS and visual quality comparisons of different distillation strategies on the RealSR benchmark. Our ASD achieves lower LPIPS scores and produces more realistic details than SDS and VSD during the early stages of training.

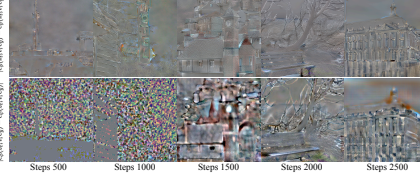


Fig. 4: Visualization of the penalty term $|\epsilon_\phi(\cdot) - \epsilon_\psi(\cdot)|$ at timestep $t = 250$. The top row shows the difference maps computed from the GT latent z_t , while the bottom row uses the synthetic latent $\hat{z}_t$ predicted by the generator.

consists of a trainable regularizer ϵ_ψ and a frozen pre-trained regularizer ϵ_ϕ . We denote the latent outputs of the distilled model as $\hat{z}_0$, and HQ latent representations as z_0 . We train G_θ by minimizing three losses: a reconstruction loss in pixel space to ensure fidelity to the ground truth, a semantic consistency loss to preserve perceptual quality, and a regularization loss derived from Adversarial Score Distillation (ASD) to improve realism. After updating the generator G_θ , we update the regularizer ϵ_ψ using the complete WGAN discriminator loss with a stabilized penalty term. After training, only the generator G_θ is used during inference. In the following, we will provide a detailed explanation of all the losses involved in the training phase.

Our Adversarial Score Distillation. Given the superior performance and stability of Adversarial Score Distillation (ASD) over VSD in image editing and text-to-3D tasks, we explore transferring its full WGAN discriminator optimization mechanism to the Real-ISR domain. We therefore adapt ASD into a new adversarial distillation framework designed specifically for Real-ISR.

Generator Optimization. The discriminator in ASD [44] is formulated as $D(x_t, y) = \log p(y_\phi|x_t) - \log p(y_\psi|x_t)$. Here, $p(y_\phi|x_t)$ and $p(y_\psi|x_t)$ represent the probabilities of classifying the noise sample $x_t = \alpha_t x_0 + \sigma_t \epsilon$ under two prompt conditions. y_ϕ serves as the prompt for real samples, while y_ψ serves as the prompt for generated samples. When x_t aligns well with the samples described by prompt y_ϕ , discriminator $D(x_t, y)$ produces a high confidence score. In contrast, it will yield a low confidence score when x_t is closer to the samples associated with prompt y_ψ .

We assume that diffusion models trained on large-scale datasets approximate real-world distributions well. Thus, we freeze a pre-trained model ϵ_ϕ as a stable proxy for real data, while introducing a learnable model ϵ_ψ , initialized from ϵ_ϕ , to track the evolving distribution of the generated samples. Under the WGAN framework conditioned on y , the generator loss is defined as follows:

$$\begin{aligned}
 \mathcal{L}'_G &= \mathbb{E}_{t,\epsilon}[-D(x_t^g, y)] = \mathbb{E}_{t,\epsilon}[-\log \frac{p(y_\phi|x_t^g)}{p(y_\psi|x_t^g)}] \\
 &\leq \mathbb{E}_{t,\epsilon}[-\log \frac{p(y_\phi|x_t^g)^\lambda}{p(y_\psi|x_t^g)}] := \mathcal{L}_G \quad \text{s.t.} \quad \lambda \geq 1,
 \end{aligned} \tag{5}$$

where $x_t^g = \alpha_t g(\theta, c) + \sigma_t \epsilon$ denotes the generated sample with noise. According to the Bayesian rule and score matching [4, 14], we can derive the transformation: $\nabla_{x_t} \log p(y|x_t) = -1/\sigma_t(\epsilon_{x_t, y, t} - \epsilon_{x_t, t})$. Thus, we have:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_G &= \nabla_{\theta} \mathbb{E}_{t, \epsilon} [\log p(y_{\psi}|x_t^g) - \lambda \log p(y_{\phi}|x_t^g)] \\ &= \mathbb{E}_{t, \epsilon} [\omega(t) \underbrace{(-(\epsilon_{x_t^g, y_{\psi}, t} - \epsilon_{x_t^g, t}))}_{\epsilon_{\psi}} + \lambda \underbrace{(\epsilon_{x_t^g, y_{\phi}, t} - \epsilon_{x_t^g, t}))}_{\epsilon_{\phi}}] \frac{\partial x_0^g}{\partial \theta}. \end{aligned} \quad (6)$$

For clarity, we rewrite the formulation by replacing the subscript notation with parameterized regularizers and refining the conditioning variable y into its embedding representation c_y , expressed as:

$$\nabla_{\theta} \mathcal{L}_G = \mathbb{E}_{t, \epsilon} [\omega(t) (\epsilon_{\psi}(x_t^g, t) + \lambda (\epsilon_{\phi}(x_t^g, c_y, t) - \epsilon_{\phi}(x_t^g, t)) - \epsilon_{\psi}(x_t^g, c_y, t)) \frac{\partial x_0^g}{\partial \theta}]. \quad (7)$$

Importantly, compared with the original formulation in Eq. (3), our Eq. (7) is no longer a **standard CFG objective**, although it retains a similar structure. Regularization in the image space requires repeated encoding and decoding, leading to high computational cost and reduced effectiveness [46]. Therefore, we reformulate it in the latent space to enhance both efficiency and stability. Considering that a well-trained VAE satisfies $E_{\phi}(x) = E_{\phi}(D(z)) \approx z$, we can approximately assume $E_{\phi}(\hat{x}_H) = \hat{z}$. This allows us to rewrite Eq. (7) in the latent space as:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_G(G_{\theta}(x_L), c_y) &= \nabla_{\hat{x}_H} \mathcal{L}'_G(\hat{x}_H, c_y) \cdot \frac{\partial \hat{x}_H}{\partial \theta} \\ &= \mathbb{E}_{t, \epsilon, \hat{z} = \alpha_t E_{\phi}(\hat{x}_H) + \beta_t \epsilon} \left[\omega(t) \left(\epsilon_{\psi}(\hat{z}_t, t) + \lambda (\epsilon_{\phi}(\hat{z}_t, c_y, t) - \epsilon_{\phi}(\hat{z}_t, t)) - \epsilon_{\psi}(\hat{z}_t, c_y, t) \right) \cdot \frac{\partial \hat{z}}{\partial \hat{x}_H} \cdot \frac{\partial \hat{x}_H}{\partial \theta} \right] \\ &= \mathbb{E}_{t, \epsilon, \hat{z} = \alpha_t \hat{z} + \beta_t \epsilon} \left[\omega(t) \left(\epsilon_{\psi}(\hat{z}_t, t) + \lambda (\epsilon_{\phi}(\hat{z}_t, c_y, t) - \epsilon_{\phi}(\hat{z}_t, t)) - \epsilon_{\psi}(\hat{z}_t, c_y, t) \right) \cdot \frac{\partial \hat{z}}{\partial \theta} \right]. \end{aligned} \quad (8)$$

Finally, Eq. (8) serves as the regularization loss to minimize the divergence between the generated and real data distributions in the latent space, and guiding G_{θ} toward generating perceptually realistic results.

Discriminator Optimization. The generator G_{θ} and discriminator D_{ψ} share a highly aligned long-term goal: both aim to model a distribution $q_{\theta}(\hat{x}_H)$ that approximates the real data distribution $p(x_H)$. Initialized from the same pre-trained teacher, they differ mainly in their optimization guidance. The generator G_{θ} is guided by an explicit and fixed teacher score, thus evolving steadily toward $p(x_H)$, the discriminator D_{ψ} is updated in a fundamentally different way. In SDS, discriminator optimization is completely ignored. In VSD, regularizer ϵ_{ψ} is trained to regress to the score of the current model distribution via the diffusion loss $\mathcal{L}_{diff}$. However, in early training, these target is rapidly moving and highly unreliable, making D_{ψ} prone to drift or produce misaligned gradients [51]. As evident in Fig. 3, VSD tends to produce strong visual artifacts

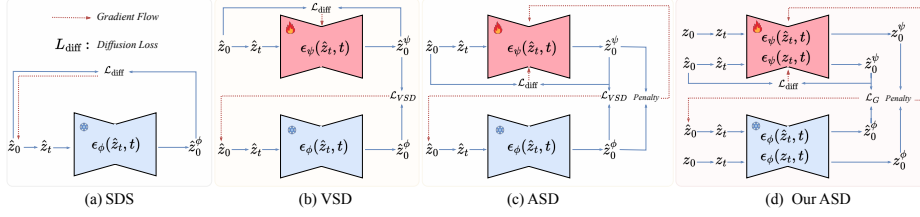


Fig. 5: Comparison of score distillation formulations and their optimization objectives. SDS ignores discriminator optimization and updates the generator solely with the diffusion loss $\mathcal{L}_{diff}$. VSD and original ASD both use the VSD objective $\mathcal{L}_{VSD}$ to guide the generator. In VSD, the discriminator ϵ_ψ is updated only using $\mathcal{L}_{diff}$. ASD further introduces an additional penalty term in the noisy latent $\hat{z}_t$. In our ASD, we adopt the complete WGAN paradigm, the generator is optimized with the adversarial loss $\mathcal{L}_G$, and the alignment penalty is compute in the GT latent z_t , leading to more stable training dynamics and finer detail reconstruction.

in the early stages of training, while SDS converges slowly in the early training phase. *Original ASD* optimizes the regularizer ϵ_ψ using the full adversarial objective as defined in Eq. (4), where both the diffusion loss and the penalty term are computed from the generator’s latent output $\hat{z}_t$. However, before convergence, the generator output $\hat{z}_t$ may deviate significantly from the real data distribution captured by the pre-trained model ϵ_ϕ , resulting in unstable penalty gradients that can hinder training or even mislead the discriminator. As shown in Fig. 4, the penalty computed with $\hat{z}_t$ contains strong noise and structural distortions during early training, providing unstable supervision. In contrast, using the GT latent z_t yields more structured and consistent signals, leading to more stable optimization of ϵ_ψ . Accordingly, we update ϵ_ψ by minimizing the following objective:

$$\nabla_\psi \mathcal{L}_D(\hat{z}_t, z_t, c_y) = \underbrace{\nabla_\psi \mathbb{E}_{t, \epsilon} [\|\epsilon_\psi(\hat{z}_t, t, c_y) - \epsilon\|^2]}_{\text{diffusion loss}} + \underbrace{\gamma \|\epsilon_\phi(z_t, t, c_y) - \epsilon_\psi(z_t, t, c_y)\|^2}_{\text{penalty term}}. \quad (9)$$

Our ASD adopts Eq. (8) as the regularization loss $\mathcal{L}_{reg}$ and optimizes the regularizer ϵ_ψ using the complete WGAN discriminator loss in Eq. (9), thereby enhancing the stability and quality of distillation. As illustrated in Fig. 3, our ASD can reconstruct fine details even in the early training stages. Fig. 5 and Table 1 present a comparison of different score distillation techniques.

Semantic Consistency Loss. Previous methods typically employ L_2 loss and LPIPS loss as reconstruction objectives, which focus on pixel-wise fidelity and perceptual similarity but often overlook high-level semantic consistency. To address this limitation, we introduce Semantic Consistency Loss (SCL), which enhances generation quality by enforcing alignment at the semantic level. SCL leverages the strong visual semantic priors of CLIP [32], pre-trained on large-scale datasets, to enforce semantic consistency between the reconstructed output and the ground truth. Given a restored image $\hat{x}_H$ and its ground truth image x_H , SCL employs the CLIP image encoder to compute their embeddings in the

Table 1: Comparison of generator and discriminator gradients.

Method	The gradient $\nabla_{\theta}\mathcal{L}_G$ used to guide the generator	The discriminator is updated using $\nabla_{\psi}\mathcal{L}_D$
VSD	$\mathbb{E}_{t,\epsilon}[\omega(t)(\epsilon_{\psi}(\hat{z}_t, t) + \lambda(\epsilon_{\phi}(\hat{z}_t, c_y, t) - \epsilon_{\phi}(\hat{z}_t, t)) - \epsilon_{\psi}(\hat{z}_t, c_y, t)) \frac{\partial \hat{z}}{\partial \theta}]$	$\nabla_{\psi}\mathbb{E}_{t,\epsilon}[\ \epsilon_{\psi}(\hat{z}_t, t, c_y) - \epsilon\ ^2]$
ASD	$\mathbb{E}_{t,\epsilon}[\omega(t)(\epsilon_{\psi}(\hat{z}_t, t) + \lambda(\epsilon_{\phi}(\hat{z}_t, c_y, t) - \epsilon_{\phi}(\hat{z}_t, t)) - \epsilon_{\psi}(\hat{z}_t, c_y, t)) \frac{\partial \hat{z}}{\partial \theta}]$	$\nabla_{\psi}\mathbb{E}_{t,\epsilon}[\ \epsilon_{\psi}(\hat{z}_t, t, c_y) - \epsilon\ ^2 + \gamma\ \epsilon_{\phi}(\hat{z}_t, t, c_y) - \epsilon_{\psi}(\hat{z}_t, t, c_y)\ ^2]$
Ours	$\mathbb{E}_{t,\epsilon}[\omega(t)(-\underbrace{(\epsilon_{\psi}(\hat{z}_t, c_y, t) - \epsilon_{\psi}(\hat{z}_t, t))}_{\epsilon_{\psi}} + \lambda\underbrace{(\epsilon_{\phi}(\hat{z}_t, c_y, t) - \epsilon_{\phi}(\hat{z}_t, t))}_{\epsilon_{\phi}}) \frac{\partial \hat{z}}{\partial \theta}]$	$\underbrace{\nabla_{\psi}\mathbb{E}_{t,\epsilon}[\ \epsilon_{\psi}(\hat{z}_t, t, c_y) - \epsilon\ ^2]}_{\text{diffusion loss}} + \underbrace{\gamma\ \epsilon_{\phi}(\hat{z}_t, t, c_y) - \epsilon_{\psi}(\hat{z}_t, t, c_y)\ ^2}_{\text{penalty on GT latent term } z_t}$

semantic space, denoted as $e_{\hat{H}}$ and e_H , respectively. These embeddings provide a high-dimensional representation of the image content in the CLIP feature space. The process is formally defined as follows:

$$e_{\hat{H}} = E_{clip}(\hat{x}_H), \quad e_H = E_{clip}(x_H), \quad (10)$$

where E_{clip} represents the CLIP image encoder. $\mathcal{L}_{SCL}$ is formulated to minimize the distance between the feature embeddings of the restored and ground truth image. The objective is formulated as follows:

$$\mathcal{L}_{SCL} = 1 - \frac{e_{\hat{H}}^{\top} e_H}{\|e_{\hat{H}}\|_2 \|e_H\|_2}. \quad (11)$$

By aligning restored and reference representations in the CLIP feature space, SCL encourages the generator to maintain semantic structure and fidelity.

Training Objective. We train the generator G_{θ} with the reconstruction loss $\mathcal{L}_{rec}$, semantic consistency loss $\mathcal{L}_{SCL}$, and regularization loss $\mathcal{L}_{reg}$. We define the reconstruction loss $\mathcal{L}_{rec}$ as the weighted sum of MSE loss and LPIPS loss:

$$\mathcal{L}_{rec}(\hat{x}_H, x_H) = \mathcal{L}_{MSE}(\hat{x}_H, x_H) + \lambda_1 \mathcal{L}_{LPIPS}(\hat{x}_H, x_H), \quad (12)$$

where λ_1 is a weight scalar. For the regularization loss $\mathcal{L}_{reg}$, we use ASD loss Eq. (8). Therefore, the overall training objective for the generator G_{θ} is:

$$\mathcal{L}(\hat{x}_H, x_H) = \mathcal{L}_{rec}(\hat{x}_H, x_H) + \lambda_2 \mathcal{L}_{SCL}(\hat{x}_H, x_H) + \lambda_3 \mathcal{L}_{reg}, \quad (13)$$

where λ_2 and λ_3 are weighting scalars. We use the complete WGAN loss Eq. (9) to update the fine-tuned regularizer ϵ_{ψ} .

4 Experiments

4.1 Experimental Settings

Datasets and Compared Methods. We train the model on the LSDIR [21] and the first 10K face images from FFHQ [16], using Real-ESRGAN [38] to synthesize LR-HR training pairs. We evaluate our method on synthetic DIV2K-Val and two real-world datasets, RealSR [2] and DrealSR [45]. We compare our model with recent advanced diffusion-based methods, which can be categorized

Table 2: Quantitative comparison against state-of-the-art methods across both synthetic and real-world datasets. The best and second best results are highlighted in **red** and **blue**, respectively.

Datasets	Metrics	Methods								
		SUPIR [53]	SeeSR [47]	ResShift [56]	HYPIR [25]	OSDiff [46]	SinSR [40]	TSD-SR [6]	InvSR [55]	Ours
DIV2K	CLIPQA $\uparrow$	0.7046	0.6865	0.5963	0.6491	0.6681	0.6488	0.7149	0.7181	0.6752
	MUSIQ $\uparrow$	64.16	68.04	60.88	65.70	67.96	62.85	70.65	66.89	68.13
	MANIQA $\uparrow$	0.5943	0.6140	0.5338	0.6119	0.6131	0.5384	0.6077	0.6424	0.6195
	PSNR $\uparrow$	22.43	23.01	24.71	22.25	23.72	24.41	22.17	22.90	24.12
	SSIM $\uparrow$	0.5479	0.6065	0.6183	0.5721	0.6109	0.6019	0.5602	0.5910	0.6210
	LPIPS $\downarrow$	0.3820	0.3469	0.3402	0.3041	0.2941	0.3240	0.2736	0.3187	0.2927
DrealSR	CLIPQA $\uparrow$	0.6737	0.6893	0.5404	0.6392	0.6958	0.6376	0.7309	0.7134	0.6885
	MUSIQ $\uparrow$	58.66	64.75	52.37	61.06	64.69	55.38	66.13	63.99	64.83
	MANIQA $\uparrow$	0.5517	0.6014	0.4748	0.6053	0.5898	0.4907	0.5820	0.5920	0.6043
	PSNR $\uparrow$	25.31	28.14	28.69	25.93	27.92	28.23	26.20	25.67	28.77
	SSIM $\uparrow$	0.6558	0.7712	0.7875	0.7197	0.7836	0.7468	0.7170	0.7131	0.7973
	LPIPS $\downarrow$	0.4122	0.3142	0.3525	0.3371	0.2968	0.3707	0.3116	0.3537	0.2896
RealSR	CLIPQA $\uparrow$	0.6619	0.6673	0.5505	0.6390	0.6687	0.6224	0.7199	0.6789	0.6598
	MUSIQ $\uparrow$	62.11	71.69	60.22	66.25	69.10	60.63	70.77	68.53	68.94
	MANIQA $\uparrow$	0.5795	0.6434	0.5402	0.6436	0.6326	0.5421	0.6311	0.6435	0.6507
	PSNR $\uparrow$	23.70	25.21	26.38	22.82	25.15	25.64	23.40	24.13	25.69
	SSIM $\uparrow$	0.6564	0.7216	0.7347	0.6783	0.7341	0.7352	0.6886	0.7125	0.7382
	LPIPS $\downarrow$	0.3650	0.3004	0.3159	0.3087	0.2921	0.3190	0.2798	0.2871	0.2805

into four groups. (1) One-step diffusion-based methods, including SinSR [40], OSDiff [46], TSD-SR [6], HYPIR [25] and InvSR [55]. (2) Multi-step diffusion-based methods, including ResShift [56], SUPIR [53], and SeeSR [47]. (3) GAN-based methods, including BSRGAN [58], Real-ESRGAN [38], FeMaSR [3], and SwinIR [22]. (4) Autoregressive (AR)-based methods, including VARSR [31] and PURE [43]. All comparison results are obtained using the officially released codes and models. Details of *GAN-based* and *AR-based methods* are provided in the supplementary material.

Evaluation Metrics. To comprehensively evaluate the performance of various methods, we adopt both full-reference and no-reference metrics. For full-reference metrics, PSNR and SSIM [42] (calculated on the Y channel in YCbCr space) are employed to assess fidelity, while LPIPS [60] quantifies perceptual quality. No-reference metrics include MUSIQ [17], MANIQA [49], and CLIPQA [36].

Implementation Details. We train our model with the AdamW optimizer [26] at a learning rate of 5e-5. The entire training process takes 12K iterations on 4 NVIDIA 5880 GPUs with a batch size of 8. The VAE encoder and diffusion network are fine-tuned using LoRA with a rank of 4, all models are initialized from the SD2.1-base model [33]. We adopt the degradation-aware prompt extraction (DAPE) [47] to extract text prompts. The weight scalars λ_1 , λ_2 , and λ_3 are set to 1, 0.2, and 1, respectively. The value of γ is fixed to -1 , and λ is set to 7.5. All experiments are conducted on the SR task with a scaling factor of 4.

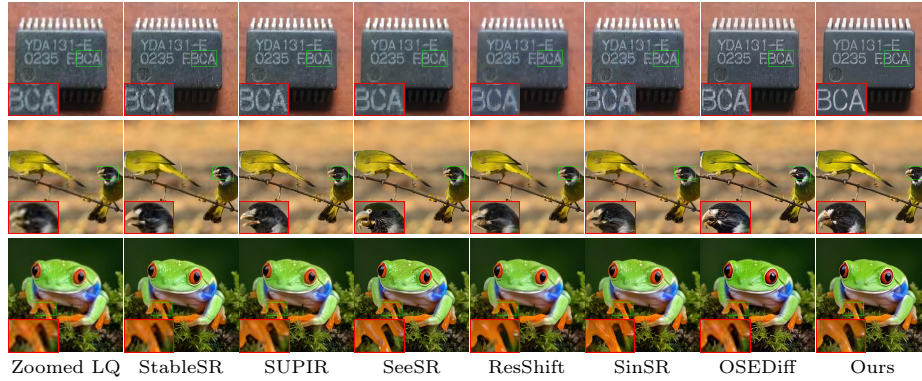


Fig. 6: Visual comparison of different methods. Please zoom in for a better view.

4.2 Comparison with State-of-the-Art Methods

Quantitative Comparisons. Table 2 presents the quantitative comparison between ASDSR and other methods. Our model demonstrates outstanding performance in PSNR, SSIM, LPIPS, and MANIQA scores, especially achieving the best results on the DrealSR dataset. ResShift achieves favorable PSNR and SSIM scores but underperforms on no-reference metrics. SinSR, distilled from ResShift, offers faster inference speed but at the cost of reduced performance. TSD-SR adopts target score matching for distillation and achieves higher CLIP-IQA and MUSIQ scores, but their relatively low PSNR and SSIM scores indicate less faithful reconstructions. In contrast, our ASDSR achieves a more balanced performance across both full-reference and no-reference metrics.

Qualitative Comparisons. Fig. 6 presents a visual comparison of different methods. As shown in the first case, ASDSR clearly stands out by faithfully reconstructing the most accurate and detailed features, while other methods produce either over-smoothed or blurred textures. A similar observation can be found in the second case, where our approach produces more natural results with significantly reduced artifacts compared to the competing methods.

Table 3: Complexity comparison among different methods. All methods are tested with an input image of size 512×512 .

	StableSR	SUPIR	SeeSR	ResShift	SinSR	InvSR	TSD-SR	Ours
Inference Step	200	50	50	15	1	1	1	1
Inference Time (s)	13.65	19.42	6.36	1.47	0.12	0.27	0.12	0.19

Complexity Comparisons. We further compare the complexity of state-of-the-art diffusion models in Table 3, focusing on inference time and the number of inference steps. All methods are evaluated on the $\times 4$ SR task using 128×128 LQ inputs, with bicubic upsampling applied before inference, on a single NVIDIA

5880 GPU. ASDSR requires only one inference steps and achieves comparable inference time. Among one-step models, our approach is only slightly slower than SinSR and TSD-SR but delivers substantially better reconstruction quality and improved stability across different degradation conditions. Note that TSD-SR reports inference time using pre-computed prompt embeddings, excluding prompt tokenization and feature extraction. After including these steps, its inference time is 0.21 seconds.

4.3 Ablation Study

Effectiveness of Semantic Consistency Loss. To validate the effectiveness of LPIPS loss [60] and Semantic Consistency Loss (SCL), we perform ablation studies by removing them separately. As shown in Table 4, removing the SCL results in a noticeable decline in LPIPS and CLIP-IQA scores. Moreover, removing the LPIPS loss leads to a significant degradation not only in LPIPS and CLIP-IQA but also in MUSIQ and MANIQA. Fig. 7 shows the visual comparisons of different losses. The absence of LPIPS or SCL loss impacts the reconstruction quality, while using both can improve detail preservation and perceptual fidelity.

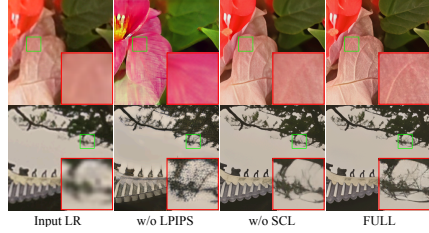


Fig. 7: Visual comparisons of different losses. Please zoom in for a better view.

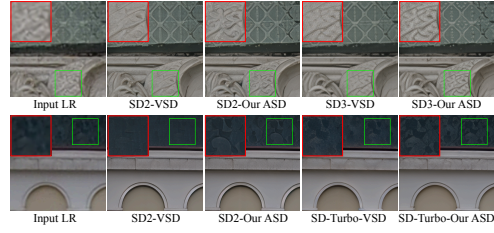


Fig. 8: Visual comparisons of VSD and our ASD across various diffusion versions.

Table 4: Comparison of different losses on the DrealSR benchmark.

Method	LPIPS ↓	MUSIQ ↑	MANIQA ↑	CLIP-IQA ↑
w/o LPIPS	0.4685	56.63	0.4663	0.5289
w/o SCL	0.2909	63.97	0.5805	0.6423
Full	0.2896	64.83	0.6043	0.6885

Table 5: Comparison of different ranks on the DrealSR benchmark.

Rank	LPIPS ↓	MUSIQ ↑	MANIQA ↑	CLIP-IQA ↑
2	0.3367	62.31	0.5690	0.6688
4	0.2896	64.83	0.6043	0.6885
8	0.2914	64.99	0.5937	0.6629

Setting of LoRA Rank. We evaluate the impact of different LoRA ranks on Real-ISR performance. As shown in Table 5, a small LoRA rank, such as 2, significantly degrades the full-reference metric LPIPS, potentially introducing more reconstruction errors. In contrast, a higher rank, such as 8, leads to lower no-reference scores. We find that setting the rank to 4 can achieve a balanced performance for both the VAE encoder and UNet.

Effect of Stable Diffusion Versions. We compare ASD and VSD across different versions of Stable Diffusion, including SD2.1-base [33], SD-turbo [34], and SD3-medium [7]. As shown in Table 6, ASD consistently outperforms VSD in terms of quality scores across all versions. Fig. 8 presents visual comparisons of the distilled results. While VSD tends to produce slightly more saturated colors and introduces structural artifacts, ASD consistently generates cleaner texture and finer structural details across all versions.

Table 6: Comparison the distillation quality between VSD and our ASD across different diffusion versions on the DrealSR benchmark.

Diffusion	Method	LPIPS ↓	MUSIQ ↑	MANIQA ↑	CLIPQA ↑
SD 2.1-Base [33]	VSD	0.3025	64.41	0.5847	0.6644
	ASD	0.2896	64.83	0.6043	0.6885
SD 3-Medium [7]	VSD	0.2993	63.90	0.5774	0.6484
	ASD	0.2687	64.85	0.5744	0.6505
SD-Turbo [34]	VSD	0.2875	63.97	0.5705	0.6572
	ASD	0.2896	63.86	0.6054	0.6728

Table 7: Comparison of distillation quality among SDS, VSD, and our ASD on the DrealSR benchmark. **Table 8:** Ablation on the weight of the semantic consistency loss on the DrealSR dataset.

Method	LPIPS ↓	MUSIQ ↑	MANIQA ↑	CLIPQA ↑	Weight	LPIPS ↓	MUSIQ ↑	MANIQA ↑	CLIPQA ↑
SDS [30]	0.2853	63.31	0.5769	0.6452	1.0	0.3025	62.18	0.5896	0.6742
VSD [41]	0.3025	64.41	0.5847	0.6644	0.5	0.3066	63.79	0.5872	0.6865
ASD	0.2896	64.83	0.6043	0.6885	0.2	0.2896	64.83	0.6043	0.6885

Effect of Adversarial Score Distillation. We further assess the superiority of ASD by comparing its distillation process with SDS and VSD. As shown in Table 7, ASD achieves the best overall perceptual quality on DrealSR, consistently outperforming SDS and VSD in terms of perceptual assessment.

Effect of CLIP Guidance Strength on Hallucination Artifacts. We conduct an ablation on the weight of the CLIP-based semantic consistency loss to examine potential issues such as color shift and hallucination artifacts. As shown in Table 8, increasing the weight consistently degrades overall performance, indicating that CLIP guidance should be kept lightweight to preserve faithful restoration. We additionally report MSSWD [11] in Table 9 to evaluate color fidelity. CLIP achieves a competitive MSSWD score that is very close to the best result, indicating that under our setting it does not noticeably compromise color fidelity while providing stronger perceptual guidance.

Comparison with Alternative Feature Extractors. We evaluate DINOv2 [28] and MAE [12] as alternative feature extractors to CLIP for the semantic consistency loss, as shown in Table 9. CLIP yields the best overall results, suggesting that its representation provides more effective semantic supervision.

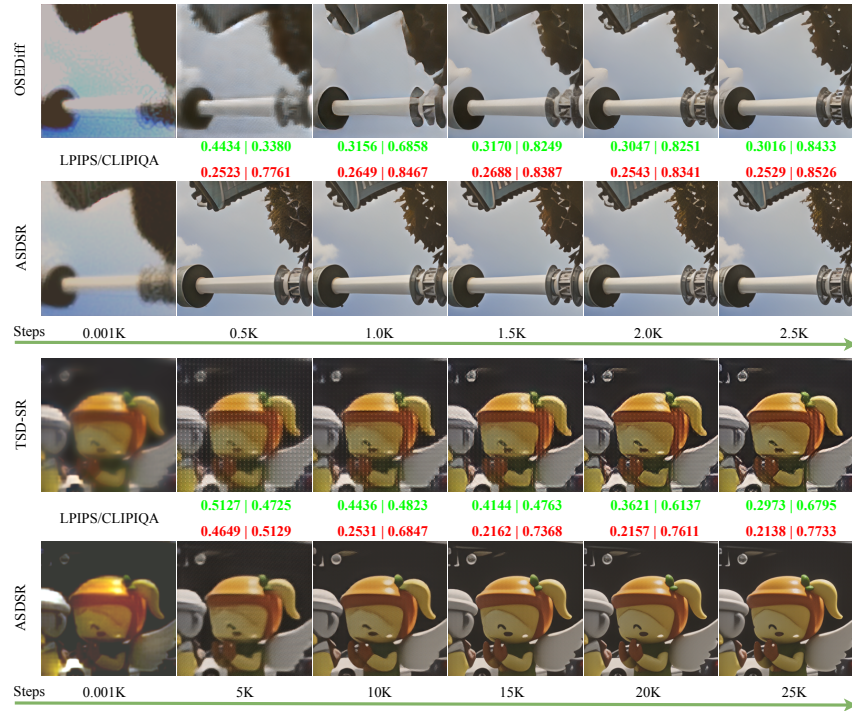


Fig. 9: Visual comparison of ASDSR with OSediff and TSD-SR at increasing training iterations, together with the corresponding LPIPS and CLIPQA scores.

Table 9: Ablation on different feature extractors on the DrealSR dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	MANIQA↑	CLIPQA↑	MSSWD↓
DINOv2 [28]	28.72	0.7896	0.2945	64.25	0.5986	0.6878	0.4472
MAE [12]	28.69	0.7810	0.2932	63.79	0.5940	0.6743	0.4492
CLIP [32]	28.77	0.7973	0.2896	64.83	0.6043	0.6885	0.4481

Effect of GT Latent in the Penalty Term. We conduct an ablation study on the choice of latent used in the penalty term in Eq. 9. As reported in Table 10, replacing $\hat{z}_t$ with z_t consistently improves reconstruction quality across both DrealSR and RealSR benchmarks, demonstrating that GT-latent replacement provides a more reliable regularization signal for robust distillation.

Comparison with OSediff and TSD-SR. We compare ASDSR with OSediff and TSD-SR throughout the training process, as illustrated in Fig. 9. For a fair comparison with TSD-SR, all experiments are conducted under the SD3-Medium framework. Under identical settings, ASDSR reaches comparable quality within only 1,000 iterations, whereas OSediff requires 2,500. Moreover, while TSD-SR produces periodic black artifacts at 25,000 iterations, ASDSR significantly mitigates these artifacts by 10,000. Overall, ASDSR converges faster and demonstrates superior stability and visual quality during training.

Table 10: Comparison between $\hat{z}_t$ and z_t latents in the penalty term.

Datasets	Penalty Latent	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	CLIPQA $\uparrow$
DrealSR	penalty on $\hat{z}_t$	28.02	0.7858	0.3132	64.41	0.5872	0.6747
	penalty on z_t	28.77	0.7973	0.2896	64.83	0.6043	0.6885
RealSR	penalty on $\hat{z}_t$	24.93	0.7134	0.3187	67.92	0.6232	0.6342
	penalty on z_t	25.69	0.7382	0.2805	68.94	0.6507	0.6598

5 Conclusion

In this work, we propose ASDSR, a one-step diffusion framework specifically tailored for Real-ISR tasks. ASDSR revisits adversarial score distillation from a WGAN perspective with a stabilized penalty term, enabling stronger and more stable gradient supervision for the generator. In addition, CLIP-derived semantic priors are introduced to further enhance semantic alignment and perceptual realism in the reconstructed images. Extensive experiments on both synthetic and real-world benchmarks verify that ASDSR consistently outperforms existing one-step and multi-step diffusion-based methods, achieving a superior balance between reconstruction fidelity, perceptual quality, and computational efficiency.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. U24A20330, 62361166670, 62276135 and 62572234, the Gusu Innovation and Entrepreneurship Leading Talent Program (Grant No. ZXL20254324), the Suzhou Key Technologies Project (Grant No. SGY2023136).

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 214–223. PMLR (06–11 Aug 2017) [3](#)
2. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV. pp. 3086–3095 (2020) [9](#)
3. Chen, C., Shi, X., Qin, Y., Li, X., Han, X., Yang, T., Guo, S.: Real-world blind super-resolution via feature matching with implicit high-resolution priors. pp. 1329–1338 (2022) [10](#)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021) [7](#)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV (2014) [1](#)
6. Dong, L., Fan, Q., Guo, Y., Wang, Z., Zhang, Q., Chen, J., Luo, Y., Zou, C.: Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In: CVPR. pp. 23174–23184 (2025) [2](#), [4](#), [10](#)

7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024) [13](#)
8. Franceschi, J.Y., Gartrell, M., Dos Santos, L., Issenhuth: Unifying gans and score-based diffusion as generative particle models. In: NeurIPS. vol. 36, pp. 59729–59760 (2023) [3](#)
9. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. ICCVW pp. 3599–3608 (2019) [1](#)
10. Gu, J., Lu, H., Zuo, W., Dong, C.: Blind super-resolution with iterative kernel correction. In: CVPR. pp. 1604–1613 (2019) [1](#)
11. He, J., Wang, Z., Wang, L., Liu, T.I., Fang, Y., Sun, Q., Ma, K.: Multiscale sliced Wasserstein distances as perceptual color difference measures. In: ECCV. pp. 1–18 (2024) [13](#)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 15979–15988 (2022) [13](#), [14](#)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020) [1](#)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021) [7](#)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022) [2](#)
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4396–4405 (2019) [9](#)
17. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV. pp. 5148–5157 (2021) [10](#)
18. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR. pp. 1646–1654 (2016) [1](#)
19. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951) [4](#)
20. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR. pp. 105–114 (2017) [1](#)
21. Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Lsdir: A large scale dataset for image restoration. In: CVPRW. pp. 1775–1787 (2023) [9](#)
22. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCVW. pp. 1833–1844 (2021) [10](#)
23. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. CVPRW pp. 1132–1140 (2017) [1](#)
24. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: ECCV. pp. 430–448 (2024) [2](#), [3](#)
25. Lin, X., Yu, F., Hu, J., You, Z., Shi, W., Ren, J.S., Gu, J., Dong, C.: Harnessing diffusion-yielded score priors for image restoration. ACM Transactions on Graphics (TOG) **44**(6), 1–21 (2025) [10](#)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019) [10](#)
27. Nguyen, T.H., Tran, A.: Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In: CVPR. pp. 7807–7816 (June 2024) [3](#)

28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) [13](#), [14](#)
29. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2024) [4](#)
30. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR (2022) [3](#), [13](#)
31. Qu, Y., Yuan, K., Hao, J., Zhao, K., Xie, Q., Sun, M., Zhou, C.: Visual autoregressive modeling for image super-resolution. In: ICML (2025) [10](#)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) [8](#), [14](#)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) [1](#), [4](#), [10](#), [13](#)
34. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: ECCV. pp. 87–103 (2024) [13](#)
35. Sun, L., Wu, R., Liang, J., Zhang, Z., Yong, H., Zhang, L.: Improving the stability and efficiency of diffusion models for content consistent super-resolution (2024) [2](#)
36. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023) [10](#)
37. Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., Guo, Y.: Unsupervised degradation representation learning for blind super-resolution. CVPR pp. 10576–10585 (2021) [1](#)
38. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCVW. pp. 1905–1914 (2021) [1](#), [4](#), [9](#), [10](#)
39. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR. pp. 606–615 (2018) [1](#)
40. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: Sinsr: diffusion-based image super-resolution in a single step. In: CVPR. pp. 25796–25805 (2024) [2](#), [10](#)
41. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. NeurIPS **36**, 8406–8441 (2023) [2](#), [3](#), [4](#), [13](#)
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004) [10](#)
43. Wei, H., Liu, S., Yuan, C., Zhang, L.: Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models. In: CVPR. pp. 18640–18650 (2025) [10](#)
44. Wei, M., Zhou, J., Sun, J., Zhang, X.: Adversarial score distillation: When score distillation meets gan. In: CVPR. pp. 8131–8141 (June 2024) [3](#), [4](#), [5](#), [6](#)
45. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: ECCV. pp. 101–117 (2020) [9](#)
46. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. In: NeurIPS. vol. 37, pp. 92529–92553 (2024) [2](#), [4](#), [7](#), [10](#)

47. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: CVPR. pp. 25456–25467 (2024) [10](#)
48. Xie, R., Zhao, C., Zhang, K., Zhang, Z., Zhou, J., Yang, J., Tai, Y.: Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation (2026) [2](#)
49. Yang, S., Wu, T., Shi, S., Gong, S., Cao, M., Wang, J., Yang, Y.: Maniq: Multi-dimension attention network for no-reference image quality assessment. CVPRW pp. 1190–1199 (2022) [10](#)
50. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: ECCV. pp. 74–91 (2023) [3](#)
51. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T.: Improved distribution matching distillation for fast image synthesis. In: NeurIPS (2024) [3](#), [7](#)
52. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: CVPR (2024) [3](#), [4](#)
53. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: CVPR. pp. 25669–25680 (2024) [2](#), [3](#), [10](#)
54. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: CVPRW. pp. 701–710 (2018) [1](#)
55. Yue, Z., Liao, K., Loy, C.C.: Arbitrary-steps image super-resolution via diffusion inversion. In: CVPR. pp. 23153–23163 (2025) [10](#)
56. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS. vol. 36, pp. 13294–13307. Curran Associates, Inc. (2023) [2](#), [10](#)
57. Zhang, A., Yue, Z., Pei, R., Ren, W., Cao, X.: Degradation-guided one-step image super-resolution with diffusion priors (2024) [4](#)
58. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV. pp. 4791–4800 (2021) [1](#), [4](#), [10](#)
59. Zhang, K., Van Gool, L., Timofte, R.: Deep unfolding network for image super-resolution. In: CVPR. pp. 3214–3223 (2020) [1](#)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018) [4](#), [10](#), [12](#)
61. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.R.: Residual dense network for image super-resolution. CVPR pp. 2472–2481 (2018) [1](#)