Spectral and Trajectory Regularization for Diffusion Transformer Super-Resolution

Jingkai Wang[1⋆], Yixin Tang[1⋆], Jue Gong[1], Jiatong Li[1], Shu Li[2], Libo Liu[2], Jianliang Lan[2], Yutong Liu[1⋆⋆], and Yulun Zhang[1⋆⋆]

[1] Shanghai Jiao Tong University, China
[2] Shenzhen Transsion Holdings Co., Ltd., China

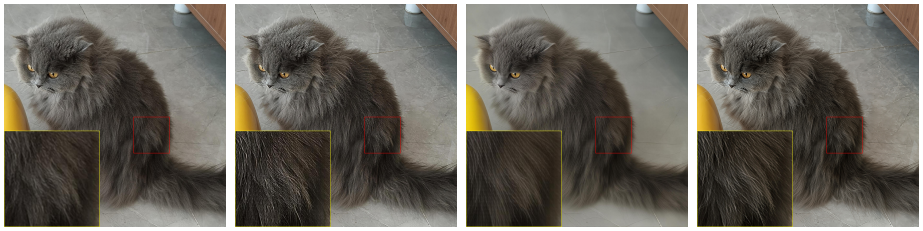
Abstract. Diffusion transformer (DiT) architectures show great potential for real-world image super-resolution (Real-ISR). However, their computationally expensive iterative sampling necessitates one-step distillation. Existing one-step distillation methods struggle with Real-ISR on DiT. They suffer from fundamental trajectory mismatch and generate severe grid-like periodic artifacts. To tackle these challenges, we propose StrSR, a novel one-step adversarial distillation framework featuring spectral and trajectory regularization. Specifically, we propose an asymmetric discriminative distillation architecture to bridge the trajectory gap. Additionally, we design a frequency distribution matching strategy to effectively suppress DiT-specific periodic artifacts caused by high-frequency spectral leakage. Extensive experiments demonstrate that StrSR achieves state-of-the-art performance in Real-ISR, across both quantitative metrics and visual perception. The code and models will be released at `https://github.com/jkwang28/StrSR`.
Keywords: Diffusion Transformer · Real-ISR · Model Distillation

1 Introduction

Real-world image super-resolution (Real-ISR) is a typical and highly practical problem in low-level vision. It originates from the classic super-resolution (SR), *i.e.*, the input low-resolution (LR) image is upscaled to a high-resolution (HR) image, with only Bicubic degradation [12, 13, 16, 86]. However, in real-world scenarios, the degradation is much more complex and unknown, making Real-ISR a highly challenging and ill-posed problem [33, 63, 83].

Substantial efforts have been devoted to Real-ISR, with a particular focus on generative approaches like GAN-based methods [29, 63, 64, 83]. These methods continuously improve and deliver steadily better performance. With the emergence and rapid development of diffusion models [19, 49, 51], Real-ISR reaches a new stage. Recent studies [35, 71, 79] increasingly adopt diffusion models for Real-ISR, leveraging their strong generative capability and scalability. Very recently, extensive research [22] has highlighted the scaling laws of diffusion models. These studies demonstrate that diffusion transformer (DiT) [48] architectures significantly outperform traditional UNets as model size and data volume grow. Given this superior capacity and predictable scaling behavior, adopting DiT for image SR is an essential and inevitable trend for the future of low-level vision.

⋆ Equal contribution.
⋆⋆ Corresponding authors: Yutong Liu and Yulun Zhang.



(a) Input **(b)** TSD-SR [17] **(c)** FluxSR [30] **(d)** StrSR (ours)

Fig. 1: For a self-captured cat image with dense fur, TSD-SR [17] and FluxSR [30] suffer from heavy grid-like artifacts, whereas our StrSR restores realistic details.

However, generating high-quality images with diffusion models usually requires dozens or hundreds of steps, which is computationally expensive. To tackle this issue, many works [17, 59, 78, 80] distill pretrained text-to-image (T2I) models for image-to-image tasks. They use either score distillation [17, 69], or adversarial diffusion distillation [31, 36], and achieve significant performance.

Despite achieving promising results, we observe that current methods still suffer when combining DiT and one-step distillation. As shown in Fig. 1, those methods [17, 30] suffer from severe artifacts and inferior performance. To analyze the reason, there is a fundamental trajectory mismatch. It is difficult to quickly and stably align the new generation trajectory (from LR to HR images) with the pretrained trajectory (from pure noise to HR images). Furthermore, during this forced trajectory alignment, models struggle to maintain photo-realistic outputs. Forcing a single large evaluation step heavily exacerbates grid-like artifacts, a problem that is uniquely severe in DiT architectures.

To tackle these challenges, we propose spectral and trajectory regularization for diffusion transformer super-resolution (StrSR). This one-step adversarial distillation framework unlocks photo-realistic generation while alleviating the severe periodic artifacts inherent to DiT models. Our contributions are as follows:

1. We propose **asymmetric discriminative distillation** to bridge the gap between multi-step and one-step generation. To bypass heavy progressive distillation and prevent model collapse, we employ a pretrained CLIP-ConvNeXt as a lightweight, texture-sensitive discriminator. Combined with a relativistic average GAN loss and approximated R1 regularization, this asymmetric design ensures stable training and superior semantic-aligned texture recovery.
2. We design **frequency distribution matching** for artifact suppression. To address the grid-like periodic artifacts arising from high-frequency spectral leakage in DiT patches, we introduce the frequency distribution loss (FDL). By minimizing the sliced Wasserstein distance between the predicted and target features across both amplitude and phase components, FDL provides a spectral constraint that effectively suppresses these artifacts.
3. We develop a **comprehensive dual-encoder architecture** that seamlessly integrates spatial adversarial distillation and spectral domain optimization. Empowered by this unified framework, StrSR achieves state-of-the-art (SOTA) performance in one-step Real-ISR. Extensive experiments confirm its superiority across both quantitative metrics and visual quality.

2 Related Works

2.1 Diffusion-based Real-ISR

Recently, the success in pretrained diffusion models [5, 49, 51, 67] enables researchers to leverage pretrained weights for downstream tasks like real-world image super-resolution. Built upon stable diffusion models [49,51] based on UNet architecture, StableSR [61] and DiffBIR [35] inject low-resolution information into the model through controlnet-like methods. OSEDiff [69] directly uses LR image as input and adopts variational score distillation (VSD) [46, 66] to distill a pretrained diffusion model for one-step super-resolution. InvSR [80] treats SR as a diffusion inversion problem to better guide image priors in large pretrained diffusion models. HYPIR [36] harnesses adversarial diffusion distillation (ADD) to distill SD-XL [49] into one-step image SR model.

With the prevalence of diffusion transformer (DiT) [48] in the text-to-image (T2I) task, some works also explored the possibility to utilize DiT architecture in image SR. TSD-SR [17] proposes target score distillation (TSD) and a distribution-aware sampling module to handle the artifacts caused by VSD in training. DiT4SR [18] trains a DiT model from scratch based on SD3 [19] MM-DiT architecture with additional LR-residual. FluxSR [30] explores one-step DiT-based SR models in terms of flow trajectory [37, 39].

2.2 Diffusion Model Acceleration

Diffusion models have strong generative abilities. However, they suffer from high computational costs during inference. This issue limits their practical deployment.

Many methods have been proposed to speed up this process. One major category is deterministic methods using a deterministic probability flow. They aim to predict the exact output of a teacher model using fewer steps. Examples include progressive distillation [53], consistency distillation [42, 43, 57, 58], and rectified flow [39], and they are easy to train. However, they usually need more steps (*e.g.*, 8 steps) to generate high-quality results. Another category is distributional methods. Instead of predicting exact outputs, they approximate the data distribution of the teacher model. This group includes adversarial training [11,26,45,54,72] and score distillation [44, 66, 77]. Some approaches even combine both techniques [9, 55, 76].

Recently, these acceleration techniques have been applied to diffusion-based low-level tasks. These works [20, 62] directly fine-tune pretrained diffusion models using score or adversarial distillation. They achieve superior performance. Notable examples in image SR include OSEDiff [69], TSD-SR [17], and HYPIR [36].

3 Methods

3.1 Preliminaries and Motivation

Diffusion Models and Rectified Flow Continuous-time generative models, such as diffusion models [23, 51, 56], build a mapping between a target data distribution $p_0(x_0)$ and a prior noise distribution $p_1(x_1)$. This transformation is driven by an ordinary differential equation over a continuous time variable $t \in [0, 1]$:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v(x_t, t), \tag{1}$$

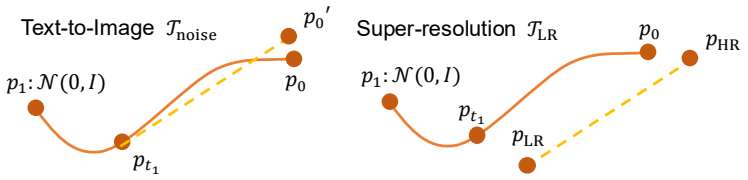


Fig. 2: Two gaps exist between the multi-step T2I model and the one-step SR model. First, a direct linear trajectory from the intermediate step $t = t_1$ to $t = 0$ cannot reach the real HR data distribution. Second, the LR data distribution mismatches the pretrained T2I model's distribution at t_1. Points in the figure represent distributions.

where $v(x_t, t)$ is the time-dependent vector field. By solving this ODE from $t = 1$ to $t = 0$, we can transport samples from noise to the data distribution.

In practice, operating directly in the high-dimensional pixel space is computationally expensive. Therefore, modern generative models project images into a continuous latent space using a pretrained variational autoencoder (VAE) [28,51] $\mathcal{E}_{\text{VAE}}, \mathcal{D}_{\text{VAE}}$. Let $z_0 = \mathcal{E}_{\text{VAE}}(x_0)$ be the latent representation of the high-quality data, and $z_1 \sim \mathcal{N}(0, \mathbf{I})$ be the standard Gaussian noise.

Rectified flow [39] connects z_0 and z_1 using a straightforward linear trajectory. The intermediate latent state z_t is defined as:

$$z_t = tz_1 + (1 - t)z_0. \tag{2}$$

By taking the derivative of z_t with respect to time t, we obtain the ground-truth vector field for this linear path:

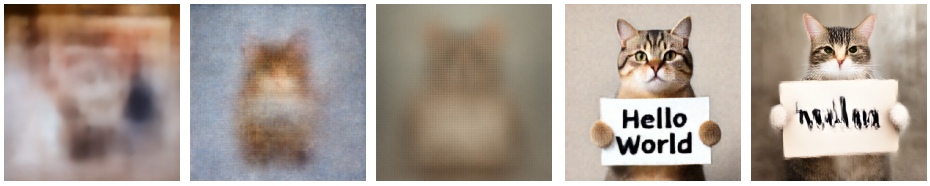
$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = z_1 - z_0. \tag{3}$$

To enable controlled generation, we introduce a condition c extracted from the input. A neural network generator $G_\theta(z_t, t, c)$ is used to predict the vector field. The network is optimized by minimizing a mean-squared error objective:

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, t, c}\left[\|G_\theta(z_t, t, c) - (z_1 - z_0)\|^2\right]. \tag{4}$$

By minimizing this loss, G_θ learns to simulate the linear mapping conditioned on c. The linear nature of rectified flow ensures a straight generation trajectory, reducing truncation errors and enabling fast inference.

Flow Matching for Real-ISR Building upon this theoretical foundation, we adapt rectified flow for image SR. Instead of synthesizing images from pure noise, we formulate the generative process as a translation from LR inputs x_{LR} (assuming $t = t_1$) to HR targets x_{HR} ($t = 0$). Operating within the compressed latent space, we redefine the initial state as the degraded latent $z_{t_1} = \mathcal{E}_{\text{VAE}}(x_{\text{LR}})$ and the target state as the clean latent $z_0 = \mathcal{E}_{\text{VAE}}(x_{\text{HR}})$. To accelerate convergence and leverage rich natural image priors, it is common practice to initialize the generator G_θ with weights from pretrained text-to-image models. Several recent methods [31,36,69] have successfully adapted these noise-pretrained weights to this deterministic $z_{t_1} \to z_0$ setting. However, as shown in Fig. 2, the distribution of degraded latents differs from intermediate state along the standard noise-to-data trajectory $\mathcal{T}_{\text{noise}}$.



SD 2.1 [51] SD 3 [19] FLUX.1 [dev] [5] Z-Image-Turbo [81] FLUX.2 [klein] [6]

Fig. 3: One-step T2I inference along a linear trajectory: UNet vs. DiT. While both models produce artifacts, the DiT-based model exhibits more severe grid-like artifacts.

This distribution shift poses a significant challenge for efficient inference. For one-step generation, where the number of function evaluations (NFE) is exactly 1, the model assumes a perfectly linear trajectory and attempts to predict the target in a single Euler step: $\hat{z}_0 = z_{t_1} - G_\theta(z_{t_1}, t_1, c)$. When image degradation is mild, the mapping distance between the pretrained noise trajectory $\mathcal{T}_{\text{noise}}$ and the new restoration trajectory $\mathcal{T}_{\text{LR}}$ is relatively small. In such cases, fine-tuning a multi-step model into a one-step model is straightforward. However, diffusion models are deployed in low-level vision primarily for their ability to restore severely degraded real-world images. Under severe degradation, the domain gap between $\mathcal{T}_{\text{noise}}$ and $\mathcal{T}_{\text{LR}}$ widens considerably. Consequently, forcing a single large evaluation step from a distant, out-of-distribution z_{t_1} leads to substantial errors, creating a distinct performance gap between multi-step and one-step generation.

Furthermore, achieving one-step generation is even more challenging with diffusion transformers (DiT). Since we fine-tune from pretrained DiT weights, our method is heavily constrained by base model. The base model is trained to generate images from pure noise. At large time steps t (near pure noise), the model primarily focuses on low-frequency global structures and ignores high-frequency details. In our one-step setting, we force a single large step directly to the target z_{HR}. As shown in Fig. 3, while taking such a large step does not cause obvious high-frequency artifacts in UNet architectures, it produces severe ones in DiT.

Overall, attempting to fine-tune a multi-step DiT model into a one-step model presents two main challenges: 1) the large mapping distance between $\mathcal{T}_{\text{noise}}$ and $\mathcal{T}_{\text{LR}}$ leads to poor results; and 2) the severe discretization error inherent in a single evaluation step worsens high-frequency artifacts in DiT architectures. To address these challenges, we need to design a novel distillation method that can effectively bridge the domain gap between multi-step and one-step trajectories while specifically mitigating DiT-related artifacts.

3.2 Model Formulation

To effectively tackle the ill-posed nature of image super-resolution, our overall architecture employs a dual-encoder design to decouple feature extraction: a vision-language model (VLM) encoder captures high-level semantic information, while a variational autoencoder extracts low-level spatial and structural representations.

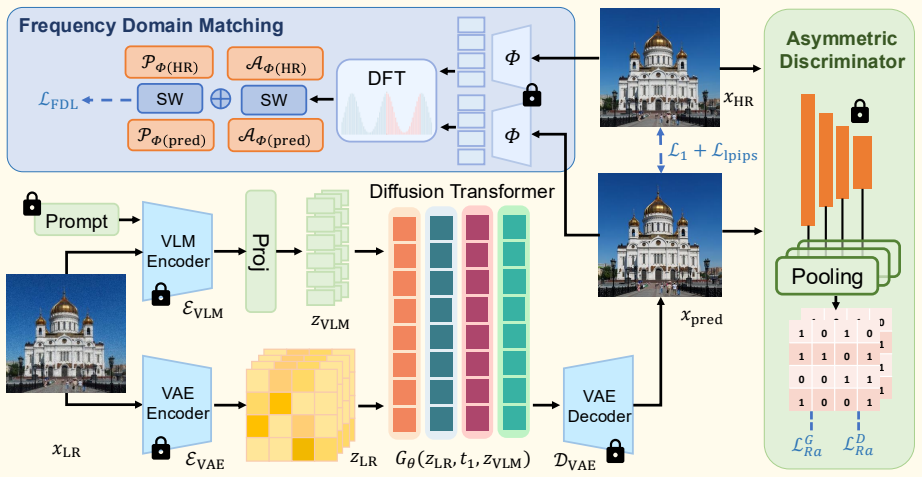


Fig. 4: Training pipeline of StrSR. The LR image x_{LR} is input into the dual-encoder $\mathcal{E}_{\text{VAE}}$ and $\mathcal{E}_{\text{VLM}}$ and following DiT G_θ to gain the predicted HR image x_{pred}. DiT is fine-tuned by LoRA. The generator and discriminator are trained alternately.

As illustrated in Fig. 4, let x_{LR} and x_{HR} denote the LR and corresponding HR images. Specifically, x_{LR} is fed into a pretrained VLM encoder, $\mathcal{E}_{\text{VLM}}$, to extract rich semantic features. To align these features with the input space of the diffusion backbone, they are processed through an adaptive MLP f_{adapt}:

$$z_{\text{VLM}} = f_{\text{adapt}}(\mathcal{E}_{\text{VLM}}(x_{\text{LR}})). \tag{5}$$

In our design, z_{VLM} serves as a direct substitute for the conventional text condition embeddings typically used in the base foundational models.

In parallel, x_{LR} is processed by a pretrained VAE encoder, $\mathcal{E}_{\text{VAE}}$, to obtain the continuous latent space representation z_{VAE}, *i.e.*, z_{LR}:

$$z_{\text{LR}} = \mathcal{E}_{\text{VAE}}(x_{\text{LR}}). \tag{6}$$

Operating within a rectified flow framework, this deterministic latent representation z_{LR} serves directly as the initial state of the generative process.

The core generation is driven by a DiT generator, G_θ. The generator takes both the spatial latent z_{VAE} and the semantic tokens z_{VLM} as inputs. These representations are formulated as token sequences and processed through the transformer blocks to predict the vector field governing the generation trajectory. The final restored high-quality image x_{pred} is obtained by decoding the integrated latent state via the VAE decoder $\mathcal{D}_{\text{VAE}}$.

3.3 Towards Photo-Realistic Restoration

Discriminative distillation is a widely-used, effective, and standard approach to align generation trajectories [31, 34, 36, 60]. Our goal is to design an efficient and stable strategy that fully exploits pretrained priors, tailored for Real-ISR.

Distribution matching-based distillation [76,77] is proven effective in generative tasks. However, when using pretrained DiTs [48] for both the generator and discriminator, we observe rapid model collapse during training. We attribute

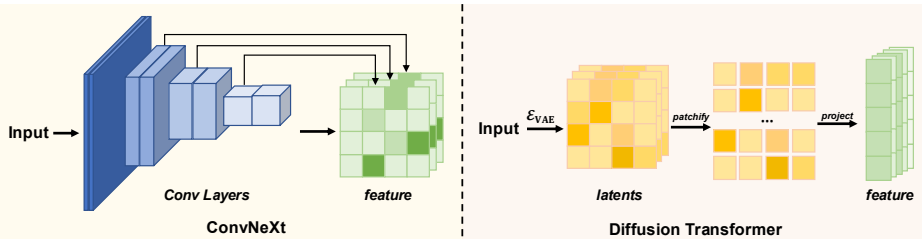


Fig. 5: Discriminator comparison. ConvNeXt's large-kernel convolutions preserve multi-scale spatial information. Conversely, DiT's patchification heavily compresses intra-patch details (edges and textures), reducing its ability to detect patch-level periodic artifacts.

this collapse to the absence of a progressive distillation [53], as highlighted in SeedVR2 [60]. To avoid complex and lengthy training pipelines while maintaining high visual quality, we discard the DiT-based discriminator. As shown in Fig. 5, we introduce a discriminator based on a pretrained CLIP-ConvNeXt [14, 36, 40], augmented with multiple trainable convolutional layers for calculating logits. As a lightweight yet powerful feature extractor, this convolutional CLIP [50] model offers three key advantages: a smaller model size, comparable semantic understanding, and superior texture perception. Most importantly, the strong local inductive bias of ConvNeXt makes it highly sensitive to high-frequency textures and grid artifacts. This asymmetric architecture perfectly mitigates the one-step DiT generator's flaw of producing local periodic artifacts.

Specifically, our discriminator adopts a PatchGAN architecture. We extract features from different stages and the final pooling layer from the CLIP-ConvNeXt backbone. These extracted features are then processed by trainable convolutional heads to produce multi-level, patch-wise logit maps for computing the GAN loss.

We adopt the relativistic average GAN (RaGAN) [25] rather than naive non-saturating GAN [21] for better visual performance and stable training convergence. The generator loss and discriminator loss are set as follows, respectively.

$$\begin{aligned} \mathcal{L}_{Ra}^{G} &= -\mathbb{E}_{z_{\text{LR}}}\left[\log g\left(x_{\text{pred}}, x_{\text{HR}}\right)\right] - \mathbb{E}_{x_{\text{HR}}}\left[\log\left(1 - g\left(x_{\text{HR}}, x_{\text{pred}}\right)\right)\right]; \\ \mathcal{L}_{Ra}^{D} &= -\mathbb{E}_{x_{\text{HR}}}\left[\log g\left(x_{\text{HR}}, x_{\text{pred}}\right)\right] - \mathbb{E}_{z_{\text{LR}}}\left[\log\left(1 - g\left(x_{\text{pred}}, x_{\text{HR}}\right)\right)\right]. \end{aligned} \tag{7}$$

Here, $x_{\text{pred}} = \mathcal{D}_{\text{VAE}}(z_{\text{LR}} - tG_\theta(z_{\text{LR}}, t, z_{\text{VLM}}))$, $g(x, y) = \varsigma(C(x) - \mathbb{E}_y[C(y)])$ with ς denoting the sigmoid function. $C(\cdot)$ is the output logits for a real or fake sample.

To better regularize the discriminator and facilitate training convergence, we also introduce approximated R1 loss [52] as:

$$L_{R1} = ||D(x_{\text{real}}) - D(\mathcal{N}(x_{\text{real}}, \sigma\mathbf{I}))||^2. \tag{8}$$

We apply Gaussian noise with small variance σ to real data when calculating discriminator loss, and adopt additional approximated R1 loss as Eq. (8). This regularization forces the discriminator to produce similar outputs for both the clean images and their noisy counterparts. Consequently, it achieves a penalizing effect similar to the standard R1 loss without the computational overhead.

In summary, our asymmetric discriminative training objectives can be:

$$\mathcal{L}_D = \mathcal{L}_{Ra}^{D} + \lambda_{R1} \cdot \mathcal{L}_{R1} \tag{9}$$

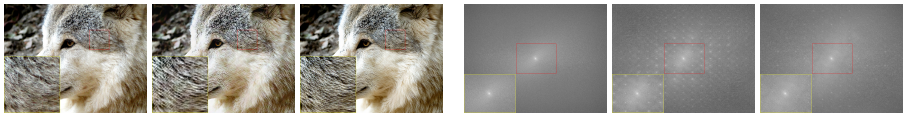


Fig. 6: Comparison of spectral decomposition with and without FDL. Without FDL, distinct periodic artifacts are visible in the left image. After applying FDL, the spectral leakage is effectively mitigated.

Type	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MAN.↑	MUS.↑	QAli.↑	Qual.↑
Multi-step Diffusion	SUPIR [79]	22.22	0.5748	0.3806	0.1660	3.4372	0.5793	63.4906	4.3862	0.6959
	DiT4SR [18]	20.76	0.5451	0.3995	0.1670	3.0084	0.6246	65.1493	4.4567	0.6987
	DiffBIR [35]	23.49	0.5837	0.3963	0.2396	4.8019	0.5730	56.7752	3.9390	0.6607
	PASD [75]	23.42	0.6229	0.4119	0.1942	4.3308	0.4657	56.1247	4.1016	0.5530
	I.R. [38]	23.35	0.6001	0.4108	0.2554	4.3449	0.5221	52.9015	3.7092	0.5486
	SeeSR [70]	23.24	0.6125	0.3473	0.1580	3.6129	0.5945	61.5973	4.4410	0.7048
One-step Diffusion	CTMSR [78]	23.93	0.6200	0.3652	0.1918	4.4132	0.4721	60.6717	4.0600	0.6872
	PiSA-SR [59]	23.36	0.6224	0.3442	0.1789	3.6403	0.5706	60.1926	4.1500	0.7109
	TSD-SR [17]	21.68	0.5549	0.3453	0.1526	2.8397	0.5454	66.5933	4.1558	0.7209
	OSEDiff [69]	23.27	0.6441	0.3972	0.2179	4.5961	0.4940	54.0388	3.9813	0.6287
	InvSR [80]	21.09	0.5979	0.4255	0.2125	4.3214	0.5027	58.3533	3.7655	0.6247
	HYPIR [36]	21.78	0.5845	0.3551	0.1457	3.3566	0.5673	63.6403	4.3359	0.7299
	SinSR [65]	22.74	0.5356	0.4630	0.2143	6.1304	0.4830	60.8577	4.1208	0.5115
	StrSR - Z	21.35	0.5798	0.3216	0.1407	3.4738	0.5727	64.6547	4.4759	0.7316
	StrSR - F	21.70	0.5982	0.2992	0.1288	3.0379	0.5864	63.6568	4.4763	0.7022

Table 1: Quantitative results (×4) on DIV2K-val dataset. I.R. denotes InstructRestore; MAN., MUS., QAli., and Qual. stand for MANIQA, MUSIQ, QAlign, and QualiCLIP, respectively. StrSR - Z/F refers to the Z-Image and FLUX implementations of StrSR.

3.4 Frequency Distribution Matching

Although the proposed asymmetric discriminative distillation can partially mitigate the artifacts caused by the mismatch between multi-step and one-step generation, notable periodic artifacts still remain, as shown in Fig. 6. Previous studies [30] have observed similar issues and analyzed them from a phenomenological perspective. They attribute these artifacts to token similarity within DiT patches, noting that the artifact period in pixel space equals the DiT patch size multiplied by the VAE scale factor. We conduct a deeper investigation into this phenomenon, showing that the fundamental cause ultimately lies in the high-frequency domain, *i.e.*, the spectral leakage problem.

To mitigate the artifacts so as to enhance realistic quality, we decide to introduce frequency domain loss regularization, frequency distribution loss (FDL) [47], to narrow the gap between predicted and real data distribution in the frequency domain. FDL was first proposed to offset the spatially misaligned data problem in image transformation tasks. The formulation and introduction of frequency in FDL provides a natural constraint to the generated distribution, resulting in anti-periodic artifact effects and better global perception.

Firstly, FDL transforms predicted and target images into feature space using a pretrained feature extractor Φ, then employs Sliced Wasserstein (SW) distance to measure the distribution distance between frequency components of the features

Type	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MAN.↑	MUS.↑	QAli.↑	Qual.↑
Multi-step Diffusion	SUPIR [79]	24.44	0.6993	0.3322	0.1982	4.5734	0.6034	64.9685	4.0683	0.6647
	DiT4SR [18]	23.24	0.6617	0.3367	0.1901	4.2814	0.6486	67.9475	3.9679	0.6592
	DiffBIR [35]	26.29	0.7242	0.3219	0.2075	5.9847	0.5851	62.4423	3.9038	0.6237
	PASD [75]	26.71	0.7615	0.2759	0.1630	4.8252	0.5416	60.9738	4.0344	0.5811
	I.R. [38]	26.10	0.7450	0.3010	0.2033	4.9780	0.5953	62.7246	3.9585	0.6122
	SeeSR [70]	26.21	0.7563	0.2786	0.1759	4.5428	0.6106	64.3097	4.0009	0.6832
One-step Diffusion	CTMSR [78]	26.18	0.7652	0.2935	0.1867	4.6381	0.5202	63.5130	3.9233	0.6750
	PiSA-SR [59]	25.60	0.7509	0.2744	0.1745	4.4039	0.6277	66.4567	3.9946	0.7073
	TSD-SR [17]	23.76	0.6984	0.2874	0.1844	3.8310	0.6139	69.0501	4.0924	0.7408
	OSEDiff [69]	24.29	0.7153	0.3440	0.2109	5.7609	0.5154	59.5873	3.8685	0.6694
	InvSR [80]	22.61	0.7114	0.3068	0.1862	4.2273	0.6332	67.1131	4.0047	0.7102
	HYPIR [36]	22.14	0.6716	0.3313	0.2004	4.1558	0.6317	68.3347	4.1322	0.7232
	SinSR [65]	26.02	0.7085	0.4014	0.2261	6.2482	0.5246	61.1541	3.8799	0.5252
	StrSR - Z	22.15	0.6606	0.3215	0.1958	4.4439	0.5941	68.5793	4.2371	0.7417
	StrSR - F	23.77	0.7236	0.2599	0.1665	3.8090	0.6333	67.3318	4.1296	0.7134

Table 2: Quantitative results (×4) on RealSR. I.R. denotes InstructRestore; MAN., MUS., QAli., and Qual. stand for MANIQA, MUSIQ, QAlign, and QualiCLIP, respectively. StrSR - Z/F refers to the Z-Image and FLUX implementations of StrSR.

after the discrete Fourier transform (DFT). The final FDL can be formulated as:

$$\mathcal{L}_{\text{FDL}}(x_{\text{HR}}, x_{\text{pred}}) = \text{SW}(\mathcal{A}_{\Phi(x_{\text{HR}})}, \mathcal{A}_{\Phi(x_{\text{pred}})}) + \lambda \cdot \text{SW}(\mathcal{P}_{\Phi(x_{\text{HR}})}, \mathcal{P}_{\Phi(x_{\text{pred}})}), \quad (10)$$

where $\mathcal{A}$ and $\mathcal{P}$ denote the amplitude and phase components of data in the frequency domain, respectively.

Taking into account the reconstruction goal in both the spatial and spectral domains, we have the following training objective for the generator:

$$\mathcal{L}_G = \mathcal{L}_1 + \lambda_1 \cdot \mathcal{L}_{\text{lpips}} + \lambda_2 \cdot \mathcal{L}_{Ra}^G + \lambda_3 \cdot \mathcal{L}_{\text{FDL}}, \quad (11)$$

where $\mathcal{L}_1$ and $\mathcal{L}_{\text{lpips}}$ represent the commonly used L1 and LPIPS losses in image restoration for perception-distortion trade-off [7].

4 Experiments

4.1 Experimental Settings

Training Datasets We choose widely-used LSDIR [32] and synthetic dataset Aesthetic-4K [82] for training. Specifically, we picked the first 60,000 images from Aesthetic-4K and all the images from LSDIR, for better generative quality and faster convergence. Images are center-cropped to 1,024×1,024 pixels during training, and the corresponding LR images are synthesized using the standard degradation pipeline proposed in RealESRGAN [63].

Evaluation Datasets We test our method on synthetic dataset DIV2K-val [2] and commonly used real-world datasets: RealSR [8] and RealLQ250 [3]. We use the same degradation pipeline as RealESRGAN [63] to generate LR images of DIV2K-val [2]. Our method is tested on full-sized images to better reflect real-world performance. All evaluations are conducted on the ×4 SR setting.

Metrics For those full-reference metrics, we evaluate pixel-level restoration using PSNR and SSIM, and perceptual metrics include LPIPS [85] and DISTS [15]. For no-reference metrics, we choose NIQE [84], MANIQA [74], MUSIQ [27], QAlign [68], and QualiCLIP [1]. All the metrics are evaluated by `pyiqa` [10].

Type	Methods	NIQE↓	MANIQA↑	MUSIQ↑	QAlign↑
Multi-step Diffusion	SUPIR [79]	3.6651	0.5785	66.1146	4.1144
	DiT4SR [18]	3.5479	0.6318	70.9467	4.2268
	DiffBIR [35]	5.0983	0.5877	66.4842	3.9950
	PASD [75]	4.5019	0.5152	61.6403	3.9688
	InstructRestore [38]	4.9263	0.5448	62.9681	3.5949
	SeeSR [70]	4.4334	0.5916	68.5174	4.1372
One-step Diffusion	CTMSR [78]	4.5791	0.5082	67.9592	3.7108
	PiSA-SR [59]	3.9149	0.6053	69.4146	4.2053
	TSD-SR [17]	3.4879	0.5829	71.0490	4.1694
	OSEDiff [69]	4.8754	0.5070	61.2067	3.7763
	InvSR [80]	4.4034	0.5719	64.7987	3.9033
	HYPIR [36]	3.9651	0.6051	69.7083	4.1993
	SinSR [65]	5.7901	0.5156	65.9640	3.7436
	StrSR - Z-Image	4.1019	0.5774	70.1176	4.2595
	StrSR - FLUX	3.4693	0.6138	69.2050	4.2851

Table 3: Comparison (×4) of different methods on RealLQ250.

Implementation Details We adopt Z-Image-Turbo [81] and FLUX.2 [klein] 4B [6] as base generator models. The Z-Image-Turbo model has 6B parameters, and it is an 8-step model distilled from Z-Image. The FLUX.2 [klein] 4B model is distilled from FLUX.2 [klein] Base 9B, and it has 4 NFEs. We instantiate the reasoning vision-language module with Qwen3-VL-4B-Instruct [4]. As both generator backbones are built on the Qwen3 [73] text encoder, Qwen3-VL provides compatible tokenization and hidden representations, allowing us to freeze the VLM and learn only a lightweight adaptive projection layer.

The discriminator's backbone is pretrained CLIP-ConvNeXt-XXLarge [14, 40, 50]. We fine-tune the pretrained generator with LoRA [24] rank=256. The generator and discriminator are updated at a 1:1 ratio. The two-stage training is performed in bf16 precision for 60,000 steps on 4 NVIDIA A800 GPUs with batch size=4. For each training stage, we use AdamW optimizer [41] with default betas, and set the learning rates of generator and discriminator to 5×10^{-6} and 1×10^{-6} respectively. The first stage takes 40,000 steps, and then another 20,000 steps are used to fine-tune with FDL loss. For the generator, we set the hyperparameters as λ_1=3, λ_2=0.1, λ_3=0 in the first stage and λ_3=0.002 in the second stage, respectively. For the discriminator, we set λ_{R1}=10. For approximated R1, we add Gaussian perturbation to real inputs with σ=0.005.

4.2 Comparison and Analysis

Compared Methods We compare our method with several state-of-the-art diffusion-based image super-resolution methods. For those representative multi-step diffusion methods, we compare with SUPIR [79], DIT4SR [18], DiffBIR [35], PASD [75], InstructRestore [38] and SeeSR [70]. As for one-step diffusion methods, our comparison includes CTMSR [78], PiSA-SR [59], TSD-SR [17], OSEDiff [69], InvSR [80], HYPIR [36], and SinSR [65].

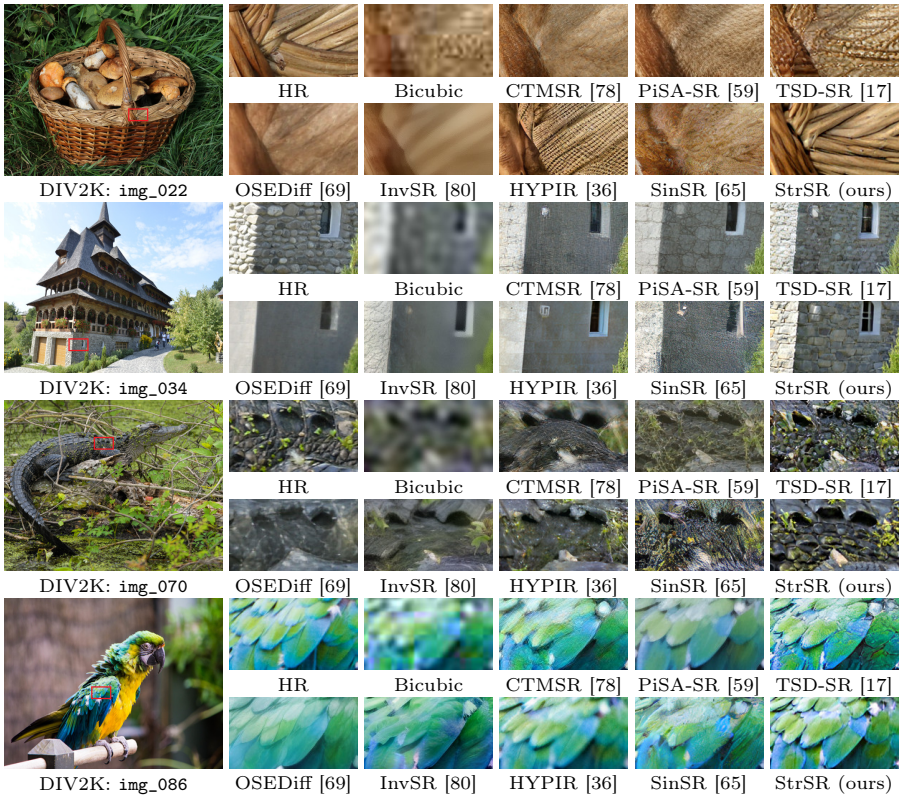


Fig. 7: Visual comparison for image SR (×4) in DIV2K-val dataset.

Quantitative Results As shown in Tabs. 1, 2, and 3, our methods achieve excellent overall performance. We compare StrSR on two different base models. The best and second-best results are colored in red and blue, respectively.

Regarding the perceptual metrics LPIPS and DISTS, our two models achieve state-of-the-art (SOTA) results across all three datasets when compared with all one-step methods. Furthermore, on the DIV2K dataset, our models even outperform all multi-step diffusion methods. For non-reference metrics, our models also show strong competitiveness. Specifically, among all one-step methods, our FLUX model achieves the best scores in NIQE and MANIQA across the datasets. This demonstrates its ability to generate highly realistic images. Meanwhile, our Z-Image model performs exceptionally well in aesthetic and visual alignment metrics. It consistently ranks in the top two for MUSIQ and QAlign.

These comparisons indicate that our proposed StrSR can capture more realistic local details and global structural information compared with previous methods. It significantly improves the overall perceptual and aesthetic quality while maintaining efficient one-step generation.

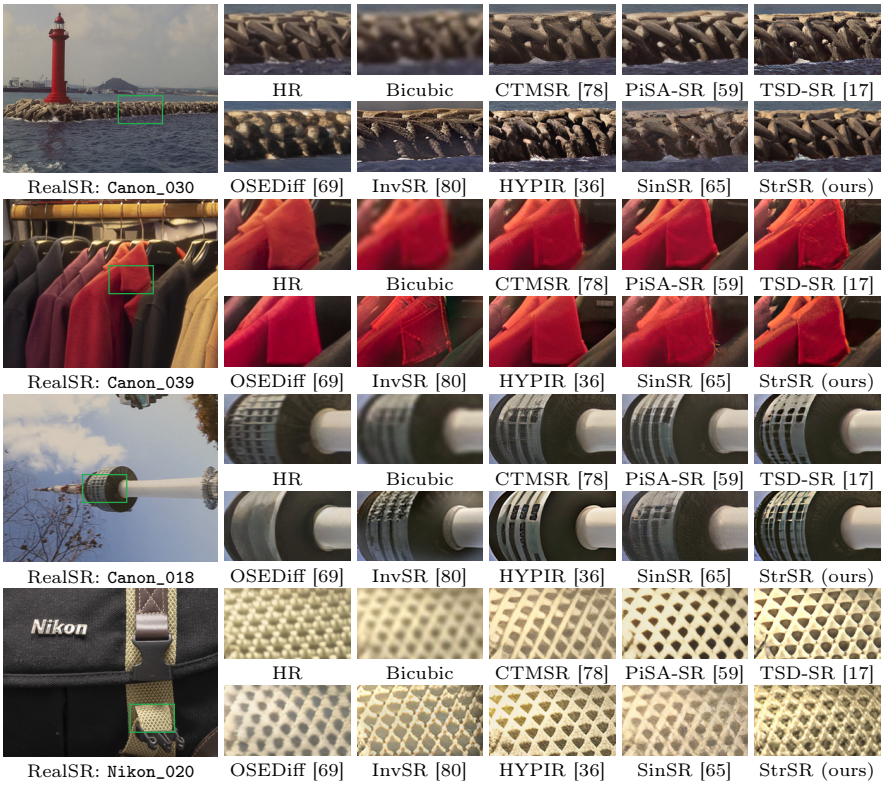


Fig. 8: Visual comparison for image SR (×4) in RealSR dataset.

Qualitative Results Visual comparisons in Figs. 7, 8, and 9 show that StrSR yields superior visual results. We use FLUX-based StrSR for visual comparisons.

For the synthetic DIV2K-val dataset, StrSR correctly recognizes high-level semantics to generate accurate texture categories. Furthermore, it produces natural and photo-realistic fine-grained details on dense high-frequency textures, such as brick walls, scales, and feathers. Additionally, while both StrSR and TSD-SR are DiT-based, StrSR avoids generating repetitive dot-like artifacts in dense high-frequency regions (e.g., basket edges) that persist in TSD-SR.

StrSR also performs excellently on the real-world datasets RealSR and RealLQ250. It segments clothing edges accurately, as seen in `Canon_039` from RealSR. Given a large gap between the LR input and HR image (*e.g.*, `Nikon_020` in RealSR), StrSR leverages semantic information to generate realistic belt textures. In contrast, other methods simply amplify the triangular artifacts present in the LR image. Finally, in image `241` from the RealLQ250 dataset, only our method successfully restores the semantic information of the dewdrops. Other methods fail: PiSA-SR generates meaningless white dots, while CTMSR and HYPIR treat the dewdrops as noise and completely smooth them out.

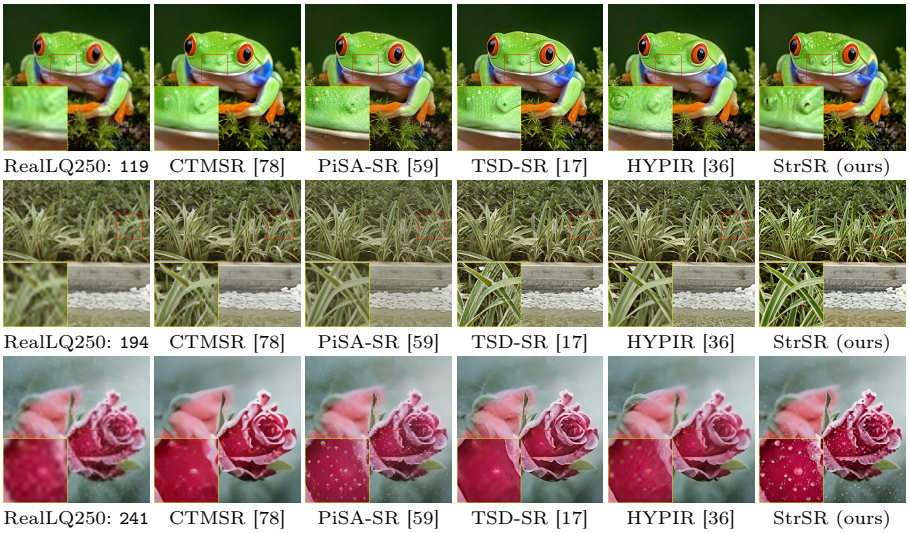


Fig. 9: Visual comparison for image SR (×4) in RealLQ250 dataset. The first column is the bicubic algorithm. Please zoom in for a better view.

4.3 Ablation Study

Frequency Distribution Matching. We extract features from the final layer of the DiT to visualize the impact of the FDL. As shown in Fig. 10, FDL effectively reduces the similarity between adjacent tokens, which suppresses the grid-like artifacts in the feature maps and promotes more natural details. Additionally, visual results in Fig. 11 demonstrate the effectiveness of frequency distribution matching in handling challenging high-frequency textures. These results indicate that FDL effectively mitigates the periodic artifacts common in DiT-based super-resolution models, enabling the generation of realistic fine details.

Asymmetric Discriminative Distillation We evaluate the effects of various GAN settings in our model's training pipeline, including the type of GAN we used (RaGAN, non-saturated GAN, and without GAN). As shown in Tab. 4, our method with RaGAN outperforms other ablation settings on multiple aesthetic and semantic metrics. What's more, Fig. 11 illustrates the visual results of different training methods, demonstrating better convergence and generative quality reached with our final GAN setting. In other words, the use of RaGAN relativistic GAN loss guarantees more stable gradient flow for the generator to learn the target data distribution in the diffusion process.

VLM Dual Encoder. Specifically, we evaluate the model under three conditions: without any text embedding (w/o emb), with a naive text encoder (w/o VLM), and with our VLM-encoded embeddings (w/ VLM). As reported in Tab. 4, integrating the VLM yields substantial improvements across all perceptual metrics, including a notable increase in MUSIQ from 62.093 to 65.106. As supported by the visual

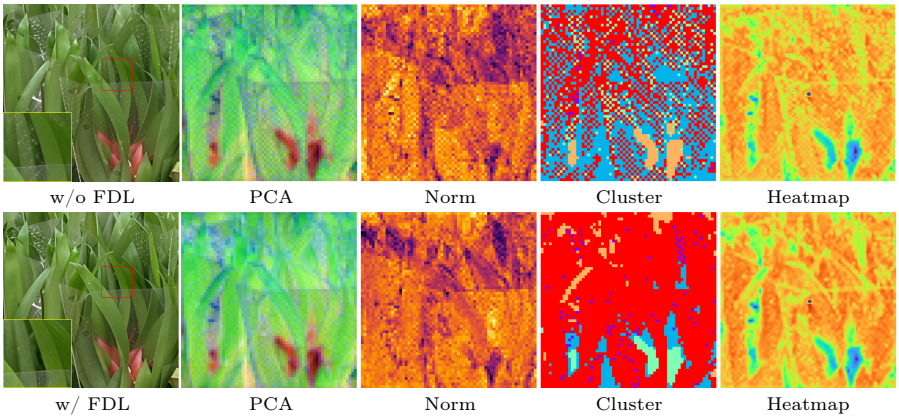


Fig. 10: Visual comparison for extracted feature maps. We extract the features of the generator's last layer for demonstration, including four widely-used visualization methods. Our FDL loss effectively alleviates artifacts in latent level.

evidence in Fig. 11, without accurate semantic conditioning (w/o emb and w/o VLM), the model struggles to comprehend complex structures, leading to severe distortions. It is evident that the VLM effectively translates LR inputs into rich semantic priors, guiding the model to generate more semantically faithful details.

4.4 Running Time

We further evaluate the inference speed for 1,024×1,024 input and compare it with various SR methods on an NVIDIA RTX A6000 GPU, as shown in Fig. 12. The results show that one-step distillation significantly improves inference speed. Remarkably, compared to other one-step models, StrSR maintains a similar speed despite using a much larger model (4B for FLUX.2 or 6B for Z-Image).

4.5 Limitations

One-step distillation significantly reduces the sampling steps compared to multi-step diffusion methods. However, our base models, Z-Image-Turbo and FLUX.2 [klein], contain far more parameters than widely used Stable Diffusion models (<1B), which limits their practicality for on-device deployment. Although StrSR achieves strong perceptual quality, its PSNR/SSIM scores, *i.e.*, pixel-level fidelity, still leave room for improvement. In addition, our current pipeline is less effective in text-rich regions, as it does not explicitly take text as input.

4.6 Conclusion

In this paper, we propose StrSR, a novel one-step adversarial distillation framework for Real-ISR. We observe that applying existing one-step distillation methods directly to DiT causes severe grid-like artifacts. This issue primarily arises

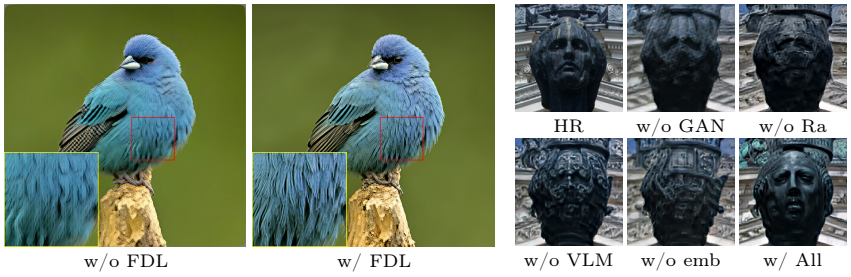


Fig. 11: (*Left*) Comparison of image textures with and without frequency distribution matching. FDL guides the model to generate more photo-realistic fine details. (*Right*) Visual comparison of the ablation results for the GAN and VLM components. StrSR accurately captures both semantic information and texture details.

Methods	MAN.	MUS.	QAli.
w/o GAN	0.4720	54.086	4.1629
w/o Ra	0.4930	61.820	4.4570
w/o emb	0.5043	58.868	4.1291
w/o VLM	0.5300	62.093	4.3435
w/ All	**0.5754**	**65.106**	**4.4779**

Table 4: Ablation study of the GAN and VLM modules on DIV2K-val. Our full model achieves the best scores, restoring semantically faithful and photo-realistic details.

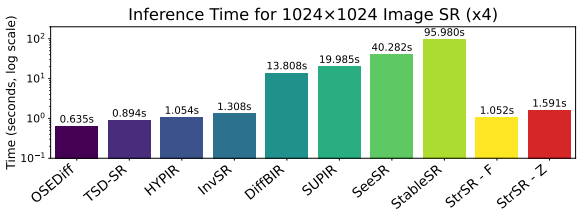


Fig. 12: Comparison of inference times for image super-resolution (×4) across various methods. The vertical axis is displayed on a logarithmic scale. Our StrSR, although utilizing a significantly larger backbone, achieves comparable inference speeds.

from generation trajectory mismatch and high-frequency spectral leakage. To address these problems, we introduce an asymmetric discriminative distillation architecture. By employing a pretrained CLIP-ConvNeXt as the discriminator, our method ensures stable training and accurate texture recovery. Furthermore, we design frequency distribution matching to map the spectral domain, which effectively suppresses the periodic artifacts. By combining these designs in a dual-encoder architecture, StrSR achieves SOTA performance in extensive experiments. Our method successfully unlocks the generative potential of DiT architectures, enabling highly efficient and photo-realistic Real-ISR in one inference step.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62501386), CCF-Tencent Rhino-Bird Open Research Fund, and CAAI-Tencent Rhino-Bird Open Research Fund. This work is also sponsored by AI Hundred Schools Program and is carried out using the Ascend AI technology stack. This work is supported in part by NSFC grant 62302299 and Huawei Explore X funds. This work is also supported by SGLabZZKT2025-09, and Oceanic Interdisciplinary Program of Shanghai Jiao Tong University, grant number SL2023ZD203.

References

1. Agnolucci, L., Galteri, L., Bertini, M.: Quality-aware image-text alignment for opinion-unaware image quality assessment. arXiv preprint arXiv:2403.11176 (2024)
2. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: CVPR (2017)
3. Ai, Y., Zhou, X., Huang, H., Han, X., Chen, Z., You, Q., Yang, H.: DreamClear: High-capacity real-world image restoration with privacy-safe dataset curation. In: NeurIPS (2024)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Black Forest Labs: Flux. `https://github.com/black-forest-labs/flux`
6. Black Forest Labs: Flux.2-klein-4b. `https://huggingface.co/black-forest-labs/FLUX.2-klein-4B`
7. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR (2018)
8. 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 (2019)
9. Chadebec, C., Tasar, O., Benaroche, E., Aubin, B.: Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In: AAAI (2025)
10. Chen, C., Mo, J.: IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: `https://github.com/chaofengc/IQA-PyTorch`
11. Chen, D.Y., Bandyopadhyay, H., Zou, K., Song, Y.Z.: Nitrofusion: High-fidelity single-step diffusion through dynamic adversarial training. In: CVPR (2025)
12. Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., Dong, C.: HAT: Hybrid attention transformer for image restoration. IEEE TPAMI (2025)
13. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: ICCV (2023)
14. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023)
15. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE TPAMI (2020)
16. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE TPAMI (2016)
17. 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 (2025)
18. Duan, Z.P., Zhang, J., Jin, X., Zhang, Z., Xiong, Z., Zou, D., Ren, J.S., Guo, C., Li, C.: DiT4SR: Taming diffusion transformer for real-world image super-resolution. In: ICCV (2025)
19. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)

20. Gong, J., Yang, T., Wang, J., Chen, Z., Liu, X., Gu, H., Liu, Y., Zhang, Y., Yang, X.: HAODiff: Human-aware one-step diffusion via dual-prompt guidance. In: NeurIPS (2025)
21. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
22. Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T.B., Dhariwal, P., Gray, S., et al.: Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701 (2020)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
25. Jolicoeur-Martineau, A.: The Relativistic Discriminator: A Key Element Missing from Standard GAN. In: ICLR (2019)
26. Kang, M., Zhang, R., Barnes, C., Paris, S., Kwak, S., Park, J., Shechtman, E., Zhu, J.Y., Park, T.: Distilling diffusion models into conditional gans. In: ECCV (2024)
27. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale Image Quality Transformer . In: ICCV (2021)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
29. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
30. Li, J., Cao, J., Guo, Y., Li, W., Zhang, Y.: One diffusion step to real-world super-resolution via flow trajectory distillation. In: ICML (2025)
31. Li, J., Cao, J., Zou, Z., Su, X., Yuan, X., Zhang, Y., Guo, Y., Yang, X.: Unleashing the power of one-step diffusion based image super-resolution via a large-scale diffusion discriminator. In: NeurIPS (2025)
32. Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., et al.: LSDIR: A large scale dataset for image restoration. In: CVPR (2023)
33. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: ICCVW (2021)
34. Lin, S., Xia, X., Ren, Y., Yang, C., Xiao, X., Jiang, L.: Diffusion adversarial post-training for one-step video generation. In: ICML (2025)
35. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Ouyang, W., Qiao, Y., Dong, C.: DiffBIR: Towards blind image restoration with generative diffusion prior. In: ECCV (2024)
36. 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. In: SIGGRAPH Asia (2025)
37. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
38. Liu, S., Ma, J., Sun, L., Kong, X., Zhang, L.: InstructRestore: Region-customized image restoration with human instructions. In: NeurIPS (2025)
39. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
40. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR (2022)
41. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
42. Lu, C., Song, Y.: Simplifying, stabilizing and scaling continuous-time consistency models. In: ICLR (2025)

43. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
44. Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., Zhang, Z.: Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In: NeurIPS (2023)
45. Luo, Y., Chen, X., Qu, X., Hu, T., Tang, J.: You only sample once: Taming one-step text-to-image synthesis by self-cooperative diffusion gans. In: ICLR (2025)
46. Nguyen, T.H., Tran, A.: Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In: CVPR (2025)
47. Ni, Z., Wu, J., Wang, Z., Yang, W., Wang, H., Ma, L.: Misalignment-robust frequency distribution loss for image transformation. In: CVPR (2024)
48. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
49. 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. arXiv preprint arXiv:2307.01952 (2023)
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
52. Roth, K., Lucchi, A., Nowozin, S., Hofmann, T.: Stabilizing training of generative adversarial networks through regularization. In: NeurIPS (2017)
53. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: ICLR (2022)
54. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast high-resolution image synthesis with latent adversarial diffusion distillation. In: SIGGRAPH Asia (2024)
55. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: ECCV (2024)
56. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
57. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. In: ICLR (2024)
58. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML (2023)
59. Sun, L., Wu, R., Ma, Z., Liu, S., Yi, Q., Zhang, L.: Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In: CVPR (2025)
60. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., et al.: SeedVR2: One-step video restoration via diffusion adversarial post-training. In: ICLR (2026)
61. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. IJCV (2024)
62. Wang, J., Gong, J., Zhang, L., Chen, Z., Liu, X., Gu, H., Liu, Y., Zhang, Y., Yang, X.: OSDFace: One-step diffusion model for face restoration. In: CVPR (2025)
63. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: ICCVW (2021)
64. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: ECCVW (2018)

65. 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 (2024)
66. 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. In: NeurIPS (2023)
67. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
68. Wu, H., Zhang, Z., Zhang, W., Chen, C., Li, C., Liao, L., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-Align: Teaching lmms for visual scoring via discrete text-defined levels. In: ICML (2024)
69. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. In: NeurIPS (2024)
70. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: SeeSR: Towards semantics-aware real-world image super-resolution. In: CVPR (2024)
71. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: DiffIR: Efficient diffusion model for image restoration. In: ICCV (2023)
72. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. In: CVPR (2024)
73. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
74. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: CVPRW (2022)
75. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: ECCV (2024)
76. 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)
77. 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)
78. You, W., Zhang, M., Zhang, L., Zhou, X., Shi, K., Gu, S.: Consistency trajectory matching for one-step generative super-resolution. In: ICCV (2025)
79. 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 (2024)
80. Yue, Z., Liao, K., Loy, C.C.: Arbitrary-steps image super-resolution via diffusion inversion. In: CVPR (2025)
81. Z-Image Team: Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)

82. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: CVPR (2025)
83. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV (2021)
84. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE TIP (2015)
85. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
86. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018)