TEASR: Training-Efficient Any-Step Diffusion Transformer for Real-World Image Super-Resolution

Xiang Gao, Chenxin Zhu, Yushun Fang,
Qiang Hu*, Xiaoyun Zhang

Shanghai Jiao Tong University, China
{frxg0918, Zcx-sjtu, ethanfang, qiang.hu, xiaoyun.zhang}@sjtu.edu.cn

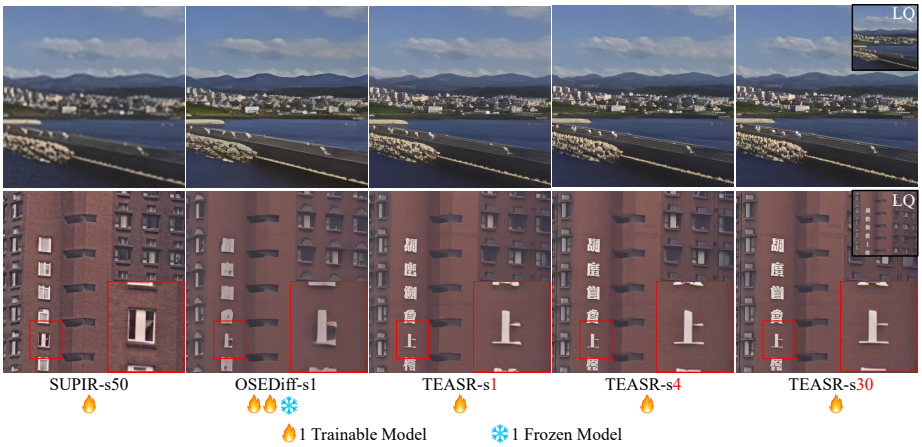

Fig. 1: Qualitative comparison of our any-step TEASR against purely multi-step and one-step method. 's' denotes the sampling steps. Red represents any-step choice. While multi-step methods require iterative refinement and single-step approaches suffer from heavy training overhead, our approach achieves satisfying quality at one-step and delivers even more superior performance as inference steps increasing.

Abstract. Diffusion models excel in Real-World Image Super-Resolution (Real-ISR) due to their powerful generative priors but suffer from slow iterative sampling. Although existing one-step distillation methods accelerate inference, they typically require auxiliary teacher models that inflate training memory and restrict scalability to large-scale architectures. Furthermore, these fixed-step models lack the flexibility to trade off speed for quality. In this paper, we propose TEASR, a training-efficient any-step diffusion framework for Real-ISR that enables both one-step and multi-step restoration within a unified model. Our key idea is to perform self-adversarial distillation within a single diffusion model, eliminating the need for auxiliary teachers or discriminators. Specifically, we propose

* Corresponding author.

a timestep-aware rectification strategy that stabilizes one-step generation across noise levels. These two designs further enables the distillation of 20B-parameter diffusion models on a single GPU, significantly improving training efficiency. Moreover, we introduce a dual-branch diffusion transformer with decoupled timestep condition to separate the current noise state and the denoising target to enhance sampling quality. Extensive experiments demonstrate that TEASR supports seamless any-step sampling and consistently outperforms state-of-the-art methods across multiple datasets. Code: `https://github.com/MediaX-SJTU/TEASR`.

Keywords: Image super-resolution · One-step Diffusion · Effective training

1 Introduction

Recently diffusion models [12, 17, 26, 28, 31, 38, 39] have achieved remarkable success in image generation and various downstream vision tasks [7, 8, 23, 47, 48]. In particular, diffusion-based approaches have established new state-of-the-art performance in Real-ISR by generating highly realistic details. However, their iterative sampling process [6, 21, 34, 44] significantly hinders real-time deployment.

To accelerate diffusion models, recent researches [9, 29, 32] have focused on advanced samplers and few-step distillation. Despite their success, these methods still face a steep quality drop-off in extreme one-step sampling scenarios. Emerging one-step distillation techniques, such as Distribution Matching Distillation(DMD) [43] and adversarial distillation training [20, 30], have achieved satisfying one-step results. Recent diffusion-based SR methods have adopt some distillation methods to achieve one-step sampling. As shown in Fig. 1, multi-step super-resolution models(e.g., SUPIR [44]) exhibit training efficiency and stability, as they require only one single trainable model, but they cannot sample with few-step even one-step. One-step super-resolution models such as OSEDiff [40] can infer quickly, but they usually maintain multiple models during training, raising memory consumption. Besides one-step methods cannot adjust the number of inference steps, abandoning the ability to increase sampling steps for more natural results. But our TEASR can perform any-step sampling, ranging from one-step to typically 30-step without extra training. As the inference steps increase, our method generates more natural restoration results.

To conclude, current one-step SR methods introduce two critical challenges in both training and application. Firstly, these methods suffer from prohibitive training overhead because they necessitate maintaining multiple extra teacher models or discriminators to provide distillation signals, which significantly limits its training efficiency and restricts their application to large-scale architectures. Secondly, these one-step models are typically confined to a fixed, preset sampling paradigm, thereby locking the restoration quality to a single forward pass and eliminating the ability to adjust inference costs according to practical needs.

To address these challenges, we propose **TEASR** (**T**raining-**E**fficient **A**ny-**S**tep Diffusion Transformer for Real-World Image **S**uper-**R**esolution), a mini-

malist framework that enables training-efficient and any-step Real-ISR under a unified conditional flow-matching formulation. TEASR performs Self-Adversarial Distillation (SAD) entirely within a single model, eliminating auxiliary teacher models or discriminators and substantially reducing training overhead, which makes it feasible to fine-tune large-scale backbones (e.g., the 20B-parameter Qwen-Image [38]) on a single NVIDIA A100 GPU. Within SAD, we propose a simple but effective Time-Aware Rectification(TAR) mechanism which adjusts the rectification strength according to the noise level, thus providing more precise signals to better align the restored results with real data distribution.

Our framework introduces a dual-branch design for DiT architecture with Decoupled Timestep Condition (DTC) that explicitly separates the generative prior from conditional guidance. In traditional self-consistent distillation [9, 32], a second timestep is introduced to represent the flow destination. By making the noise branch conditioned on the current timestep t_{cur} and the condition branch conditioned on the target timestep t_{tar}, TEASR preserves fine-grained noise-level information in self-consistent objective, enabling the model to precisely reason about the instantaneous noise state and the intended destination jointly without information loss. This design eliminates information entanglement, thereby ensuring superior restoration fidelity in any-step sampling, providing a practical adjustment to trade inference efficiency for restoration quality.

To summarize, our contributions are as follows:

- We propose TEASR, a training-efficient any-step SR method that utilizes Self-Adversarial Distillation(SAD) to eliminate the need for auxiliary teachers or discriminators during training. As a core refinement within this framework, we introduce Time-Aware Rectification(TAR) to adaptively weight the distillation signals, ensuring the restoration results are more precisely aligned with the real data distribution.
- We introduce a dual-branch DiT architecture with Decoupled Timestep Condition(DTC). This design decouples the two timesteps from the self-consistent objective, assigning them separately to the noise and condition branches to ensure that both timestep information is fully preserved within the attention blocks to achieve more faithful reconstruction in any-step sampling.
- Extensive experiments demonstrate that TEASR supports seamless any-step sampling and consistently outperforms state-of-the-art methods across multiple datasets. Our any-step sampling have good performance scalability, enabling a boost in restoration quality as the number of inference steps increases.

2 Related Work

2.1 Real-World Image Super-Resolution

Real-world image super-resolution(Real-ISR) aims to recover high-quality(HQ) images from low-quality(LQ) inputs suffering from unknown and complex degra-

dations. Early approaches predominantly relied on pixel-wise regression objectives (e.g., PSNR-oriented methods). While effective for full-reference pixel-wise metric scores, these methods often produce over-smoothed results lacking perceptual realism due to the ill-posed nature of the problem. To address this, Generative Adversarial Networks(GANs) [10, 35] were introduced to model the data distribution, yielding visually pleasing textures. GAN-based method such as Real-ESRGAN [35] has made significant progress. However, GANs frequently suffer from training instability, mode collapse, and artifacts, limiting their reliability and robustness in practical applications.

Recently, diffusion models [12, 17, 26, 28, 38] have emerged as the new state-of-the-art in Real-ISR, surpassing GANs in both training stability and generative modeling ability, achieving visually more realistic performance. Methods like [6, 21, 34, 44] leverage powerful pretrained diffusion priors to handle real-world degradations effectively. More recently, diffusion transformers(DiTs) [26] have shown great potential in scaling capabilities of diffusion models. DiT4SR [6] adopts pretrained large-scale DiT models and finetune them to Real-ISR task, achieving superior perceptual realism in restoration results.

2.2 Diffusion Distillation

The iterative sampling process of diffusion models poses a significant bottleneck for real-time applications. Consequently, extensive research has focused on diffusion distillation to reduce sampling steps. Diffusion distillation aims to learn a direct mapping from noise to data (or between arbitrary time steps). Progressive distillation [29] adopts a pre-trained teacher model to progressively distill a student model to fewer steps. Consistency Models(CMs) [9, 32] are optimized by enforcing self-consistency along the probability flow trajectory. While these methods successfully reduce sampling to few steps (e.g., 4–10 steps), they often struggle to maintain high quality when distilled to one-step.

To achieve true one-step generation, recent works have explored distribution matching distillation(DMD)-like and adversarial distillation. Methods such as DMD [43] and Adversarial Diffusion Distillation(ADD) [30] introduce auxiliary discriminators or teacher models to align the student's one-step output distribution with the real data distribution. Similarly, in the SR domain, methods like OSEDiff [40] and HYPIR [22] have adopted multi-model distillation strategies to achieve one-step super-resolution. However, two critical limitations persist: these state-of-the-art one-step frameworks typically require maintaining multiple co-resident models (e.g., teacher, student, and discriminator) during training. This significantly increases GPU memory consumption (often requiring about 2–3$\times$ the memory of a single model), making it prohibitively expensive to distill large-scale foundation models like Qwen-Image or Flux [17]. Furthermore, most existing one-step SR methods [5, 7, 40] are fixed to a specific one-step sampling paradigm, losing the ability to leverage additional computation for higher quality when latency permits.

While recent works [3, 27] have demonstrated the potential of self-adversarial distillation within a unified model for text-to-image generation, directly adapting

this paradigm to the Real-ISR task is highly non-trivial. Unlike text-to-image generation, Real-ISR strictly demands both reconstruction fidelity and perceptual realism. Furthermore, precise perception of noise levels is also essential for Real-ISR to avoid over-smoothing while effectively removing degradations [7]. To handle this domain shift and address the open challenge of enabling any-step controllability, TEASR significantly advances the unified distillation framework for Real-ISR.

3 Method

Consistency Models(CMs) [32] aim to enforce self-consistency across the Probability Flow ODE (PF-ODE) trajectory. Advanced frameworks [9] generalize this property by introducing a target timestep t_{tar}(here we use s for simplicity), which acts as an explicit destination for each sampling step. This formulation defines a velocity-based consistency function $v_\theta(\mathbf{z}_t, t, s)$ that predicts the transition from current state z_t at time t to a targeted state at time s. The core objective is to ensure that a direct step from t to s remains consistent with an iterative path passing through an intermediate state. Formally, let t, s be the current and target timesteps respectively, and $m = \frac{t+s}{2}$ be the midpoint. The generalized self-consistent objective is defined as:

$$\mathbf{v}_1 = v_\theta(z_t, t, s) \tag{1}$$

$$z_m = z_t - \frac{t-s}{2} v_\theta(z_t, t, m) \tag{2}$$

$$\mathbf{v}_2 = \frac{1}{2} v_\theta(z_t, t, m) + \frac{1}{2} v_\theta(z_m, m, s) \tag{3}$$

$$L_{\text{cons}} = ||\mathbf{v}_1 - \mathbf{v}_2||_2 \tag{4}$$

where $\mathbf{v}_1$ represents the single-step shortcut, and $\mathbf{v}_2$ denotes the integrated velocity of the two-step composition. In this paradigm, s is no longer a passive parameter but a crucial guidance signal that anchors the denoising direction. Without the explicit conditioning on s, the model would suffer from temporal ambiguity, failing to distinguish between trajectories with different refinement destinations. This dual-timestep dependency (t, s) is essential for enabling any-step inference but introduces the challenge of interference or entanglement between two temporal signals, which we will address in Sec. 3.3.

Framework overview. Fig. 2 illustrates the overall Self-Adversarial Distillation architecture of TEASR. Our framework operates within a unified conditional flow matching paradigm and introduces a Self-Adversarial Distillation scheme that eliminates the need for external teacher or discriminator models. The core design features a dual-branch DiT backbone with Decoupled Timestep Condition, where the noise branch is conditioned on the current timestep(t_{cur}) and the condition branch is conditioned on the target timestep(t_{tar}). Besides we propose a Time-Aware Rectification to precisely provide distillation signals to better straighten the generation path.

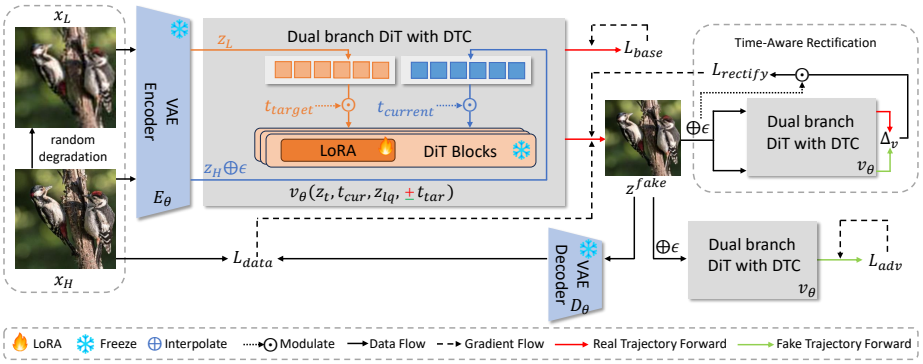

Fig. 2: Overall SAD Framework. We propose a dual-branch DiT with DTC to introduce LQ condition and decouple two timesteps. For training, we first optimize the real trajectory with flow matching loss and self-consistent loss(L_{base}). We then estimate a fake image within one-step and decode it back to pixel domain to calculate its reconstruction loss(L_{data}). Then the fake trajectory learns the one-step generated image data distribution(L_{adv}). Finally we perform TAR with both real and fake trajectory to optimize the generation process of the one-step estimated image(L_{rectify}).

3.1 Self-Adversarial Distillation for SR

Twin trajectories for Self-Adversarial Distillation. Inspired by recent works [3, 27], we introduce twin trajectories in one model to create a self-adversarial signal, thus eliminating the need for extra models.

Firstly, we train the model to serve as both an instantaneous velocity predictor and a self-consistent path-integral predictor by minimizing a standard flow-matching objective together with a self-consistent regularizer:

$$L_{\text{base}} = L_{\text{flow}} + \lambda_{\text{cons}} L_{\text{cons}} \tag{5}$$

where L_{flow} enforces accurate instantaneous velocity estimation, and L_{cons} encourages consistency between the one-forward(single-step) trajectory and the multi-forward(multi-step) trajectory. Accurate instantaneous velocity prediction serves as a stable anchor for distillation training and provides the exact real score required by the final one-step DMD stage(see Sec. 3.2). In parallel, the self-consistent path-integral component aligns single-forward and multi-forward trajectories, supporting robust few-step sampling and facilitating subsequent one-step distillation.

Next, we obtain a one-step fake estimate by directly predicting the clean data from pure noise. Specifically, we sample $\epsilon \sim \mathcal{N}(0, I)$, and perform a single velocity prediction conditioned on the target timestep t_{tar} as follows

$$z^{\text{fake}} = \epsilon - v_\theta(\epsilon, t, z_{\text{lq}}, t_{\text{tar}}), \tag{6}$$

where $t = 1, t_{\text{tar}} = 0$ and z_{lq} denotes the encoded LQ image. To model the distribution of z^{fake}, we define a fake trajectory by corrupting the stopped-gradient

z^{fake} with Gaussian noise. Concretely, we sample $\epsilon^{\text{fake}} \sim \mathcal{N}(0, I)$ and $t' \sim \mathcal{U}(0, 1)$, and form

$$z_{t'}^{\text{fake}} = t'\,\epsilon^{\text{fake}} + (1 - t')\,\text{sg}(z^{\text{fake}}), \tag{7}$$

$$L_{\text{adv}}(\theta) = \mathbb{E}_{z_0, z_{\text{lq}}} \left\| v_\theta\big(z_{t'}^{\text{fake}}, t', z_{\text{lq}}, -t'\big) - \big(\epsilon^{\text{fake}} - \text{sg}(z^{\text{fake}})\big) \right\|_2, \tag{8}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. We optimize the fake trajectory solely with the flow-matching objective by making it conditioned on the corresponding current timestep and fake target timestep, implemented as $-t'$.

Finally, the learned fake trajectory provides a fake score that can be contrasted with the real score from the real trajectory to perform DMD-based rectification. Concretely, we re-corrupt the one-step fake estimate z^{fake} with gaussian noise and feed the perturbed latent into both trajectories to obtain paired real/fake scores. Their discrepancy yields a rectification signal that pulls the one-step prediction toward the real data distribution while pushing it away from the fake estimate distribution:

$$z_s^{\text{fake}} = s\,\epsilon^{\text{rectify}} + (1 - s)\,\text{sg}(z^{\text{fake}}), \tag{9}$$

$$\Delta_v\big(z_s^{\text{fake}}\big) = v_\theta\big(z_s^{\text{fake}}, s, z_{\text{lq}}, -s\big) - v_\theta\big(z_s^{\text{fake}}, s, z_{\text{lq}}, s\big), \tag{10}$$

$$L_{\text{rectify}}(\theta) = \mathbb{E}_{z_0, z_{\text{lq}}} \left\| v_\theta(z_t, t, z_{\text{lq}}, t_{\text{tar}}) - \text{sg}\Big[v_\theta(z_t, t, z_{\text{lq}}, t_{\text{tar}}) - \Delta_v\big(z_s^{\text{fake}}\big) \Big] \right\|_2 \tag{11}$$

where $t = 1, t_{\text{tar}} = 0$, $s \sim \mathcal{U}(0, 1), \epsilon^{\text{rectify}} \sim \mathcal{N}(0, I)$ and $z_t = \epsilon$ given by Eq. (6). Overall, optimizing the fake-trajectory objective encourages the model to represent the distribution of the fake estimate z^{fake}, while the DMD rectification explicitly guides the one-step output z^{fake} closer to the real data distribution and away from the fake one, establishing a self-adversarial training scheme.

Pixel-space fidelity anchoring. In super-resolution, a reconstruction objective is typically imposed to ensure fidelity to the ground-truth HQ image. Moreover, at the early stage of training, the model often fails to produce a meaningful one-step estimate z^{fake}, which in turn makes the learning of the fake trajectory ineffective. To address this issue, we introduce a pixel-domain fidelity constraint on z^{fake} by decoding it into image space and directly supervising it with the HQ target. Specifically, we compute a mixed reconstruction loss consisting of MSE and LPIPS:

$$x^{\text{fake}} = \mathcal{D}(z^{\text{fake}}), \tag{12}$$

$$L_{\text{data}}(\theta) = L_{\text{MSE}}\big(x^{\text{fake}}, x_0\big) + \lambda_{\text{lpips}}\, L_{\text{lpips}}\big(x^{\text{fake}}, x_0\big). \tag{13}$$

where $\mathcal{D}(\cdot)$ denotes the VAE decoder. This pixel-domain constraint anchors the one-step prediction to the HR supervision, preventing early collapse of z^{fake} and providing a reliable training signal that stabilizes subsequent fake-trajectory modeling and DMD rectification.

3.2 Time-Aware Rectification

In the DMD stage of self-adversarial distillation, the rectification term $\Delta_v(z_s^{\text{fake}})$ represents the velocity discrepancy between the real and fake trajectories. This term serves as the primary distillation gradient signal to refine the one-step estimate z^{fake}. However, we observe that the efficacy of this rectification is inherently coupled with the noise level s. Naively sampling $s \sim \mathcal{U}(0,1)$ and assigning uniform weights fails to achieve optimal convergence, as it overlooks the varying reliability of guidance signals across different temporal stages. While prior works [5, 46] have also observed this issue, they typically resort to computationally expensive multi-forward averaging or simply fixed timesteps, which may compromise the quality of the distilled model.

Specifically, the precision of the estimated velocity discrepancy undergoes significant fluctuations as the noise level s varies:

Noise-dominant Regime ($s \to 1$): When s is large, the latent z_s^{fake} is dominated by stochastic noise, resulting in a low signal-to-noise ratio. In this region, the structural information of the noisy latent is heavily corrupted, causing the instantaneous velocities predicted by both real and fake trajectories to exhibit high variance. Consequently, the rectification signal in this regime suffers from high velocity ambiguity and prediction error, which can introduce suboptimal updates and slow convergence.

Data-dominant Regime ($s \to 0$): Conversely, as s approaches the clean data manifold, the latent state retains more identifiable structural features. This allows the model to yield more accurate and consistent velocity predictions. The resulting discrepancy Δ_v provides a high-accuracy guidance signal that is crucial for recovering sharp edges and intricate textures, thus aligning the estimate z^{fake} more closely with real data distribution.

To provide the optimization process more reliable guidance, we propose a Timestep-Aware Rectification scheduler. Instead of treating all timesteps s equally, TAR applies a time-dependent scaling factor to suppress the unreliable gradients from high-noise states and amplify the precise signals from low-noise timesteps. The rectified velocity discrepancy is reformulated as:

$$\Delta_v(z_s^{\text{fake}}) \leftarrow w(s) \cdot \Delta_v(z_s^{\text{fake}}), \quad \text{where } w(s) = \frac{(1-s)^2}{s}. \tag{14}$$

By incorporating the weight $w(s)$, the training objective prioritizes timesteps with higher guidance reliability. This strategy effectively mitigates the misguidance caused by noise-dominant signals, facilitating the generation of high-frequency details, reducing one-step artifacts while ensuring the global structural integrity of the one-step estimate z^{fake}.

3.3 Decoupled Timestep Condition

Dual branch conditioning. DiT models operate on a tokenized representation of the noisy latent, which naturally enables spatial conditioning by concatenating the condition tokens with the noisy-latent tokens. The subsequent self-attention

Algorithm 1: Any-Step Sampling of TEASR

Input: Input LQ data $\boldsymbol{x}_{\text{lq}}$, Inference steps N, VAE encoder E and Decoder D, model $\boldsymbol{v}_\theta$

Output: Restored result $\boldsymbol{x}_{\text{pred}}$

$\boldsymbol{z}_{\text{lq}} \leftarrow E(\boldsymbol{x}_{\text{lq}})$

Sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\boldsymbol{z}_{\tau_0} \leftarrow \boldsymbol{\epsilon}$

Construct time steps $\{\tau_0, \tau_1, \ldots, \tau_N\}$ linearly from 1 to 0

$\Delta\tau \leftarrow 1/N$

for $i \leftarrow 0$ **to** $N-1$ **do**

$\quad \boldsymbol{v}_{\text{pred}} \leftarrow \boldsymbol{v}_\theta(\ \underbrace{\boldsymbol{z}_{\tau_i}, \tau_i}_{\text{Noise Branch}}\ ,\ \underbrace{\boldsymbol{z}_{\text{lq}}, \tau_{i+1}}_{\text{Cond Branch}}\ )$

$\quad \boldsymbol{z}_{\tau_{i+1}} \leftarrow \boldsymbol{z}_{\tau_i} - \boldsymbol{v}_{\text{pred}} \cdot \Delta\tau$

end

$\boldsymbol{x}_{\text{pred}} \leftarrow D(\boldsymbol{z}_{\tau_N})$

return $\boldsymbol{x}_{pred}$

layers allow the noisy-latent tokens to attend to the condition tokens, thereby injecting conditional information in a spatially aligned manner. Following OminiControl [33], we adopt a prior-preserved dual-branch conditioning design: we freeze the parameters associated with the first half of the token sequence (the noise branch) to preserve the pretrained generative prior, and only finetune the second half (the condition branch) to efficiently learn SR-specific conditioning.

Decoupled Timestep Condition. Prior consistency-style methods (e.g., Shortcut, Meanflow) typically need two timesteps to represent t_{cur} and t_{tar}. To let the model be aware of these two timesteps, they introduce the t_{tar} by fusing it with the t_{cur}, either via numerical adjustment (e.g., $t \leftarrow t_{\text{cur}} - \frac{t_{\text{tar}}}{2}$) or by summing timestep embeddings (e.g., $\text{temb} \leftarrow \text{temb}_1(t_{\text{cur}}) + \text{temb}_2(t_{\text{tar}})$). Such fusion will entangle the two different timestep signals, thus losing fine-grained noise-level information and target information, which is particularly harmful for super-resolution where fidelity is highly sensitive to the diffusion noise level.

We instead decouple the two timestep conditions: the noise branch is conditioned on the current timestep t_{cur}(noise level), while the condition branch is conditioned on the target timestep t_{tar}(destination of the flow step). This branch-wise injection preserves both timestep signals explicitly through attention, allowing the model to jointly reason about the current noise state and the intended denoising target without entangling them into a single embedding.

3.4 Any-Step Sampling

The proposed TEASR framework, by virtue of its SAD, DTC and TAR, overcomes the limitations of previous fixed-step SR methods. It facilitates any-step sampling, allowing users to dynamically navigate the trade-off between inference efficiency and restoration quality at test time without retraining at different timesteps.

Table 1: Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. "Models" denotes the number of maintained models during training. The best and second best results of each metric are highlighted in red and blue, respectively.

Datasets	Methods	Step(s)	Model(s)	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	MUSIQ↑	MANIQA↑	NIQE↓
RealSR	StableSR	200	1	24.70	0.7085	0.3018	0.2288	128.51	65.78	0.6221	5.91
	DiffBIR	50	1	24.75	0.6567	0.3636	0.2312	128.99	64.98	0.6246	5.53
	SUPIR	50	1	25.03	0.6705	0.3749	0.2510	120.97	57.86	0.5656	7.79
	DiT4SR	40	1	23.50	0.6657	0.3215	0.2251	118.55	67.77	0.6564	5.98
	OSEDiff	1	3	25.15	0.7341	0.2921	0.2128	123.49	69.09	0.6335	5.65
	TSD-SR	1	3	24.81	0.7172	0.2743	0.2105	114.48	71.18	0.6345	5.12
	TEASR	1	1	24.83	0.7198	0.2542	0.2003	103.97	68.50	0.6265	5.43
DrealSR	StableSR	200	1	28.03	0.7536	0.3284	0.2269	148.98	58.51	0.5601	6.52
	DiffBIR	50	1	26.71	0.6571	0.4557	0.2748	166.79	61.07	0.5930	6.31
	SUPIR	50	1	26.69	0.6655	0.4328	0.2767	152.69	55.47	0.5339	8.88
	DiT4SR	40	1	25.40	0.6657	0.3869	0.2508	163.05	65.75	0.6287	6.92
	OSEDiff	1	3	27.92	0.7835	0.2968	0.2165	135.29	64.65	0.5895	6.49
	TSD-SR	1	3	27.77	0.7559	0.2966	0.2136	135.32	66.61	0.5850	5.93
	TEASR	1	1	28.54	0.7876	0.2722	0.2088	129.06	62.70	0.5582	6.59
DIV2K-Val	StableSR	200	1	23.26	0.5726	0.3113	0.2048	24.44	65.92	0.6192	4.76
	DiffBIR	50	1	23.64	0.5647	0.3524	0.2128	30.72	65.81	0.6210	4.70
	SUPIR	50	1	23.15	0.5436	0.3632	0.2256	27.99	62.61	0.5938	6.32
	DiT4SR	40	1	21.78	0.5485	0.3448	0.2100	30.57	68.07	0.6430	4.93
	OSEDiff	1	3	23.72	0.6109	0.2942	0.1975	26.34	67.96	0.6147	4.71
	TSD-SR	1	3	23.02	0.5808	0.2673	0.1821	23.29	71.69	0.6226	4.32
	TEASR	1	1	24.36	0.6144	0.2740	0.1865	23.40	65.72	0.5737	5.12

As detailed in Algorithm 1, the sampling procedure is formulated as a discrete denoising along the learned ODE trajectory. The core of our iterative refinement lies in the interaction between the dual-branch with the DTC design. For each step $i \in \{0, \ldots, N-1\}$, the model $\boldsymbol{v}_\theta$ performs co-reasoning by simultaneously receiving two decoupled timestep signals of t_{cur} and t_{tar}.

4 Experiments

4.1 Experimental Settings

Implementations. We train TEASR for the ×4 Real-ISR task based on Qwen-Image-2512. The training process takes ∼47,000 steps on 2 NVIDIA A100 GPUs with a batch size of 1 and a gradient accumulation steps of 2. We use the AdamW [24] optimizer with learning rate 5e-5, β =(0.9, 0.99), weight decay=0. We add LoRA [13] to the condition branch with rank and alpha of 128. The noise branch is frozen. We set gradient clipping to 1.0.

Datasets. For training, we utilize LSDIR [19] and the first 10,000 images of FFHQ [14], reaching a total of 94,991 images, where LQs are synthesized online following common degradation pipeline proposed in Real-ESRGAN [35] with the same parameter settings as previous works. For Real-ISR evaluation, we adopt two real-world datasets: RealSR [2] and DRealSR [37], and a synthetic dataset

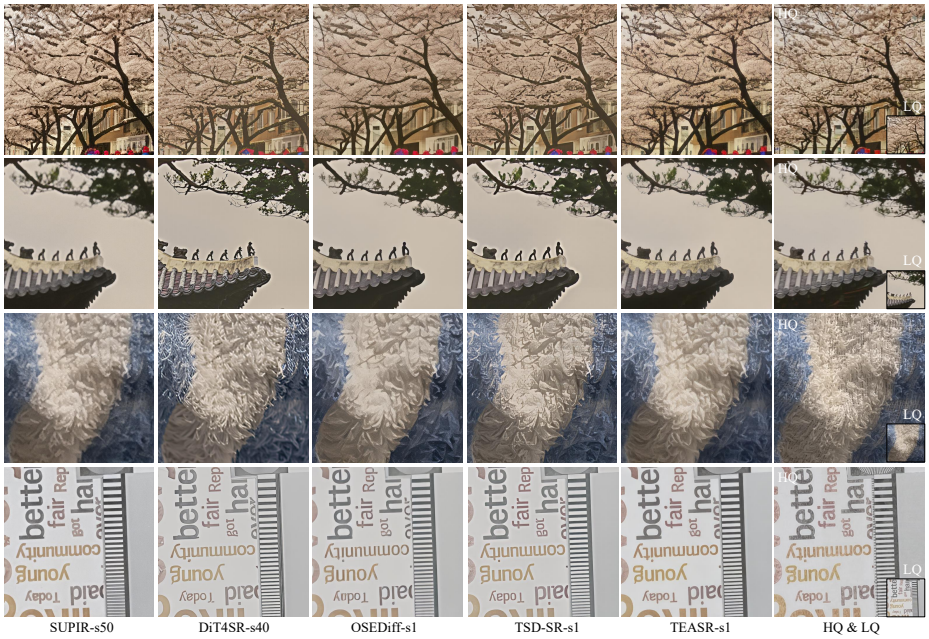

Fig. 3: Qualitative comparison with state-of-the-art methods. 's' denotes the sampling steps. Compared with other state-of-the-art methods, TEASR produces the most natural outputs and the closest match to the ground truth.

DIV2K-Val [1] following OSEDiff [40]. Without specific instruction, our model's prompt is left empty, other methods retain their default settings.

Metrics. Both full-reference(FR) and no-reference(NR) metrics are employed for evaluation. FR metrics include pixel-level metrics PSNR and SSIM [36], perceptual-level metrics LPIPS [45] and DISTS [4] and distribution-level FID [11]. NR image quality metrics include IQA methods MUSIQ [15], MANIQA [42] and NIQE [25].

Compared Methods. We compare TEASR with SOTA diffusion-based one-step methods OSEDiff [40], TSD-SR [5], multi-step methods StableSR [16], DiffBIR [21], SUPIR [44] and DiT4SR [6]. We obtain all results using their official open-source codes and models under their default settings unless otherwise specified.

4.2 Comparison with State-of-the-Arts

Quantitative Comparisons. To validate the effectiveness of our method, we compare TEASR against state-of-the-art multi-step and one-step methods on both synthetic and real-world benchmarks. As shown in Tab. 1, TEASR achieves an outstanding performance on restoration fidelity. Although OSEDiff obtains a slightly higher PSNR and SSIM on RealSR, it comes at the heavy memory consumption of maintaining three separate models during training. In contrast, our

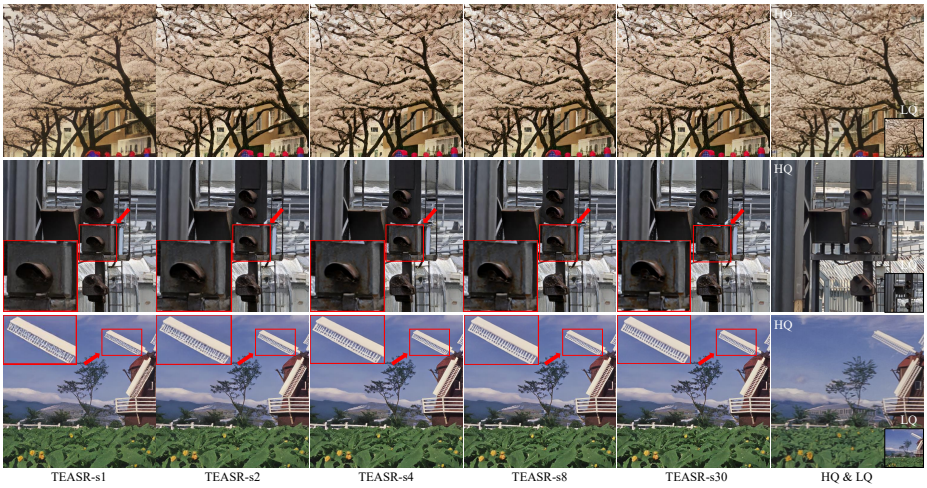

Fig. 4: Qualitative comparison of TEASR across different sampling steps. 's' denotes the diffusion sampling steps. More sampling steps result in more natural and realistic outputs. Please zoom in for a better view.

TEASR relies on only a single trainable model while securing the best perceptual metrics on RealSR, including the lowest LPIPS (**0.2542**) and FID (**103.97**). Furthermore, our method demonstrates superior generalization on DrealSR and DIV2K-Val, directly outperforming OSEDiff in PSNR (**28.54** vs. 27.92, and **24.36** vs. 23.72, respectively) and achieving the highest SSIM (**0.7876** and **0.6144**), while also demonstrating competitive performance in NR image quality metrics like NIQE. This comprehensive improvement indicates that our DTC precisely injects decoupled timestep information into the dual-branch DiT architecture, ensuring exceptional restoration fidelity without compromising the efficiency of our unified SAD framework.

Qualitative Comparisons. Fig. 3 presents the visual comparisons of our method against state-of-the-art models on the Real-ISR datasets. As observed in the first three rows, TEASR consistently provides high-quality and natural restoration results across complex natural scenes and fine-grained textures. While other methods often struggle with over-smoothing or unnatural details, our approach effectively recovers intricate visual elements with high fidelity. Furthermore, the bottom row highlights our method's precision in text and structural restoration. The result generated by TEASR exhibits well-defined stroke structures and is completely free from halo artifacts, clearly demonstrating that TEASR is highly capable of reconstructing complex characters, confirming TEASR's superior perceptual realism.

Table 2: Inference efficiency comparison.

Method	Backbone	Param.	Step(s)	Peak Mem.	Latency / Image
OSEDiff	SD2.1	0.8B	1	5934 MB	0.15 s
DiT4SR	SD3	2B	30	17244 MB	4.70 s
TEASR	Qwen-Image	20B	1 / 4 / 30	59190 MB	0.72 / 2.48 / 10.31 s

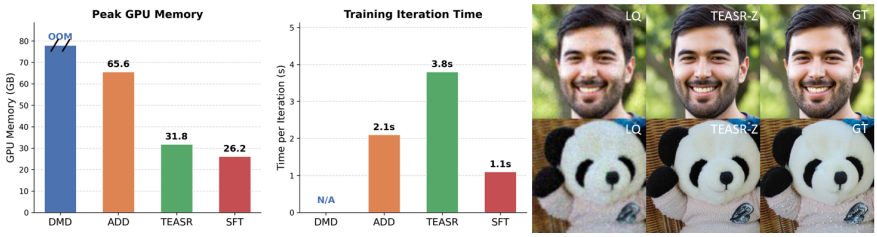

Fig. 5: Comparisons of different distillation methods on Z-Image. The left panel illustrates the training memory consumption and time per iteration, while the right panel presents qualitative results generated by TEASR-Z.

4.3 Validation of Training and Inference Efficiency

To validate the training efficiency of our proposed method, we evaluate the training VRAM consumption and time per iteration of different distillation methods on the same backbone Z-Image-6B using a single NVIDIA A100 GPU. As shown in Fig. 5, TEASR trained on Z-Image (denoted as TEASR-Z) achieves the lowest memory overhead while maintaining a reasonable time cost. Besides the right side is the visualizations of TEASR-Z, the results indicate that our framework can be applied to relatively small backbones, demonstrating our framework is robust to model scale.

Tab. 2 reports the inference efficiency comparisons on A100 GPU. Despite using the large backbone, TEASR still maintains practical inference speed while supporting one-step, few-step, multi-step SR within a unified model, achieving superior performance.

4.4 Any-Step Sampling

Different from previous one-step SR methods, TEASR is able to perform any-step sampling, achieving better quality with more sampling steps. Specifically according to Tab. 3, inference with more sampling steps results in a decrease in FR metrics while an increase in NR metrics. Visual comparison is shown in Fig. 4. In the 2nd and 3rd row of Fig. 4, increasing the number of steps will generate more details such as rust on the traffic light, make the texture of the windmill blades more straight and clear. To conclude, when increasing sampling steps, TEASR maintains good structural consistency, progressively generates more natural details and remove some unreal artifacts. More visual comparisons can be found in the **supplementary materials**.

Table 3: Quantitative comparison across different sampling steps tested on RealSR of TEASR. The best and second best results of each metric are highlighted in red and blue, respectively.

Step(s)	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	MUSIQ↑	MANIQA↑	NIQE↓
1	24.83	0.7198	0.2542	0.2003	103.97	68.50	0.6265	5.4334
2	24.02	0.6993	0.2707	0.2105	115.76	68.65	0.6188	5.2770
4	23.87	0.6966	0.2704	0.2085	113.86	68.71	0.6345	5.3127
8	23.69	0.6892	0.2849	0.2169	117.02	68.52	0.6321	5.2344
30	23.61	0.6837	0.2913	0.2183	118.41	68.42	0.6305	5.2327

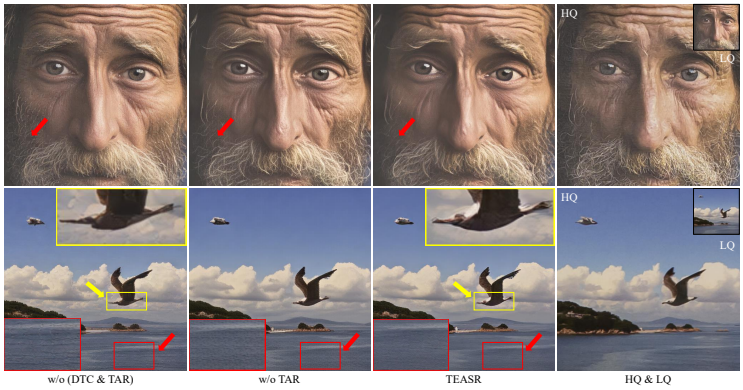

Fig. 6: Qualitative comparison of ablation study. Please zoom in for a better view.

4.5 Ablation Study

Effectiveness of SAD. Self-Adversarial Distillation is our baseline for achieving few-step even one-step sampling. Without SAD, it will degenerate into a general multi-step model. So we do not conduct ablation study on it.

Effectiveness of DTC. We explore the effectiveness of Decoupled Timestep Condition. Because DTC is a structural-level design which is not entangled with TAR(a loss-level design), we simply compare the baseline with the setting that only introduces dual-branch DiT with DTC. According to the 1st and 2nd row of Tab. 4, DTC is superior to traditional timestep embedding fusion strategy in FR metrics, indicating that DTC helps model precisely restore most of the structural content on the image. In the 1st and 2nd columns of Fig. 6, results with DTC look more similar to the HQs. We consider it is due to the complete retention of both timesteps condition information, which can more correctly provide the model the information of noise level and target destination. This phenomenon may be more obvious in Real-ISR than in t2i generation.

Effectiveness of TAR. We validate the effectiveness of Time-Aware Rectification, or, to elaborate, why is it necessary to add a time-aware weight to the rectification of the one-step generation process rather than use a constant weight. According to 2nd and 3rd row of Tab. 4, the introduce of TAR results in

Table 4: Ablations. 'DTC' denotes Decoupled Timestep Condition. 'TAR' denotes Time-Aware Rectification. All results are sampled with one-step and tested on RealSR dataset.

Structures		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	MUSIQ↑	MANIQA↑	NIQE↓
DTC	TAR								
✗	✗	24.93	0.6840	0.2904	0.2317	114.49	68.26	0.6288	4.84
✓	✗	25.52	0.7329	0.2335	0.1994	96.11	66.39	0.6252	6.72
✓	✓	24.83	0.7198	0.2542	0.2003	103.97	68.50	0.6265	5.43

a decrease in FR metrics while an increase in NR metrics. Visual differences are shown in 2nd and 3rd columns of Fig. 6, removing the Time-Aware Rectification results in the checkerboard-like artifact which is often observed in DiT architecture SR methods [18, 41]. This artifact is a DiT-specific artifact happened in distilled methods when the generation path is not sufficiently optimized. TEASR successfully removes such artifact due to our TAR design provides a better guidance for the generation process.

5 Discussions

Limitations. While TEASR significantly reduces training overhead by maintaining only a single model, optimizing multiple composite objectives within a limited model capacity remains a challenge, which may lead to slight trade-offs in restoration quality under extreme conditions. Furthermore, although TEASR provides flexibility across different sampling steps, achieving fine-grained, independent control over fidelity and realism within a fixed step remains an open question for future exploration.

Conclusions. In this work, we present TEASR, a training-efficient any-step framework for Real-ISR. By introducing SAD, we eliminate the need for maintaining extra models, enabling large-scale DiT training with minimal resource requirements, significantly improving training efficiency. Coupled with our dual-branch design with DTC and TAR, TEASR offers remarkable sampling flexibility and superior restoration performance. We hope our efficient paradigm serves as a practical baseline for future research in Real-ISR.

6 Acknowledgments

This work is supported by National Natural Science Foundation of China (62271308,62571322), STCSM(24ZR1432000, 22DZ2229005), 111 plan (BP0719010), and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3086–3095 (2019)
3. Cheng, Z., Sun, P., Li, J., Lin, T.: Twinflow: Realizing one-step generation on large models with self-adversarial flows. arXiv preprint arXiv:2512.05150 (2025)
4. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
5. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23174–23184 (2025)
6. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18948–18958 (2025)
7. Fang, Y., Chen, Y., Yin, S., Hu, Q., Yao, J., Zhang, Y., Zhang, X., Wang, Y.: One-step diffusion transformer for controllable real-world image super-resolution. arXiv preprint arXiv:2511.17138 (2025)
8. Fang, Y., Liu, L., Gao, X., Hu, Q., Cao, N., Cui, J., Chen, G., Zhang, X.: Robust id-specific face restoration via alignment learning. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 122–137. Springer (2025)
9. Frans, K., Hafner, D., Levine, S., Abbeel, P.: One step diffusion via shortcut models. In: International Conference on Learning Representations. vol. 2025, pp. 34668–34684 (2025)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. 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. Iclr **1**(2), 3 (2022)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
16. Kim, H., Choi, Y., Kim, J., Yoo, S., Uh, Y.: Exploiting spatial dimensions of latent in gan for real-time image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 852–861 (2021)

17. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
18. Li, J., Cao, J., Guo, Y., Li, W., Zhang, Y.: One diffusion step to real-world super-resolution via flow trajectory distillation. arXiv preprint arXiv:2502.01993 (2025)
19. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1775–1787 (2023)
20. Lin, S., Xia, X., Ren, Y., Yang, C., Xiao, X., Jiang, L.: Diffusion adversarial post-training for one-step video generation. arXiv preprint arXiv:2501.08316 (2025)
21. 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: European conference on computer vision. pp. 430–448. Springer (2024)
22. 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)
23. Liu, L., Cai, C., Shen, S., Liang, J., Ouyang, W., Ye, T., Mao, J., Duan, H., Yao, J., Zhang, X., et al.: Moa-vr: A mixture-of-agents system towards all-in-one video restoration. IEEE Journal of Selected Topics in Signal Processing (2025)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
25. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
26. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
27. Peng, Y., Zhu, K., Liu, Y., Wu, P., Li, H., Sun, X., Wu, F.: Facm: Flow-anchored consistency models. arXiv preprint arXiv:2507.03738 (2025)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
29. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
30. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
31. Shen, S., Liang, J., Cai, C., Geng, C., Duan, H., Zhang, X., Hu, Q., Zhai, G.: Agentic retoucher for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29114–29125 (2026)
32. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: International Conference on Machine Learning. pp. 32211–32252. PMLR (2023)
33. Tan, Z., Xue, Q., Yang, X., Liu, S., Wang, X.: Ominicontrol2: Efficient conditioning for diffusion transformers (2025)
34. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
35. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021)

36. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
37. 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: European conference on computer vision. pp. 101–117. Springer (2020)
38. 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 (2025)
39. Wu, H., Shen, S., Hu, Q., Zhang, X., Zhang, Y., Wang, Y.: Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3944–3953. IEEE (2025)
40. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems **37**, 92529–92553 (2024)
41. Wu, Z., Sun, Z., Zhou, T., Fu, B., Cong, J., Dong, Y., Zhang, H., Tang, X., Chen, M., Wei, X.: Omgsr: You only need one mid-timestep guidance for real-world image super-resolution. arXiv preprint arXiv:2508.08227 (2025)
42. 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. arXiv e-prints arXiv:2204.08958 (Apr 2022)
43. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
44. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25669–25680 (2024)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
46. Zhang, T., Duan, Z.P., Guo, C.L., Jiang, P.T., Li, B., Cheng, M.M., Li, C.: Time-aware one step diffusion network for real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30543–30552 (2026)
47. Zhang, Z., Zhang, Z., Liao, J., Meng, X., Hu, Q., Zhu, S., Zhang, X., Qin, L., Wang, W.: Lato: Landmark-tokenized diffusion transformer for fine-grained human face editing. arXiv preprint arXiv:2509.25731 (2025)
48. Zhang, Z., Gao, X., Wang, Z., Zhang, X., et al.: Td-bfr: Truncated diffusion model for efficient blind face restoration. arXiv preprint arXiv:2503.20537 (2025)