

AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation

Rui Xie¹, Chen Zhao¹, Kai Zhang¹, Zhenyu Zhang¹, Jun Zhou², Jian Yang¹, Ying Tai^{1*}

¹State Key Laboratory for Novel Software Technology, Nanjing University, Jiangsu, China.

²Southwest University, Chongqing, China.

*Corresponding author(s). E-mail(s): yingtai@nju.edu.cn;

Contributing authors: ruixie0097@gmail.com; 2518628273@qq.com; kaizhang@nju.edu.cn; zhangjesse@foxmail.com; zhouj@swu.edu.cn; csjyang@nju.edu.cn;

Abstract

Blind super-resolution methods based on Stable Diffusion (SD) demonstrate impressive generative capabilities in reconstructing clear, high-resolution (HR) images with intricate details from low-resolution (LR) inputs. However, their practical applicability is often limited by poor efficiency, as they require hundreds to thousands of sampling steps. Inspired by Adversarial Diffusion Distillation (ADD), we incorporate this approach to design a highly effective and efficient blind super-resolution method. Nonetheless, two challenges arise: First, the original ADD significantly reduces result fidelity, leading to a perception-distortion imbalance. Second, SD-based methods are sensitive to the quality of the conditioning input, while LR images often have complex degradation, which further hinders effectiveness. To address these issues, we introduce a Timestep-Adaptive ADD (TA-ADD) to mitigate the perception-distortion imbalance caused by the original ADD. Furthermore, we propose a prediction-based self-refinement strategy to estimate HR, which allows for the provision of more high-frequency information without the need for additional modules. Extensive experiments show that our method, AddSR, generates superior restoration results while being significantly faster than previous SD-based state-of-the-art models (e.g., $7\times$ faster than SeeSR).

Keywords: Super-resolution, Diffusion model, Generative prior, Efficiency

1 Introduction

Blind super-restoration (BSR) aims to convert low-resolution (LR) images that have undergone complex and unknown degradation into clear high-resolution (HR) versions. Differing from classical super-resolution [Dong et al \(2016\)](#); [Kim et al \(2016\)](#); [Ledig et al \(2017\)](#); [Chen et al \(2021\)](#); [Zhang et al \(2022a\)](#); [Tai et al \(2017a,b\)](#), where the degradation process is singular and known, BSR

is crafted to enhance real-world degraded images, imbuing them with heightened practical value.

Generative models, *e.g.* generative adversarial network (GAN) and diffusion model, have demonstrated significant superiority in BSR task to achieve realistic details. GAN-based models [Zhang et al \(2021\)](#); [Wang et al \(2021\)](#); [Xia et al \(2023b\)](#); [Liang et al \(2022\)](#); [Chen et al \(2022\)](#); [Liang et al \(2021\)](#) learn a mapping from the distribution of

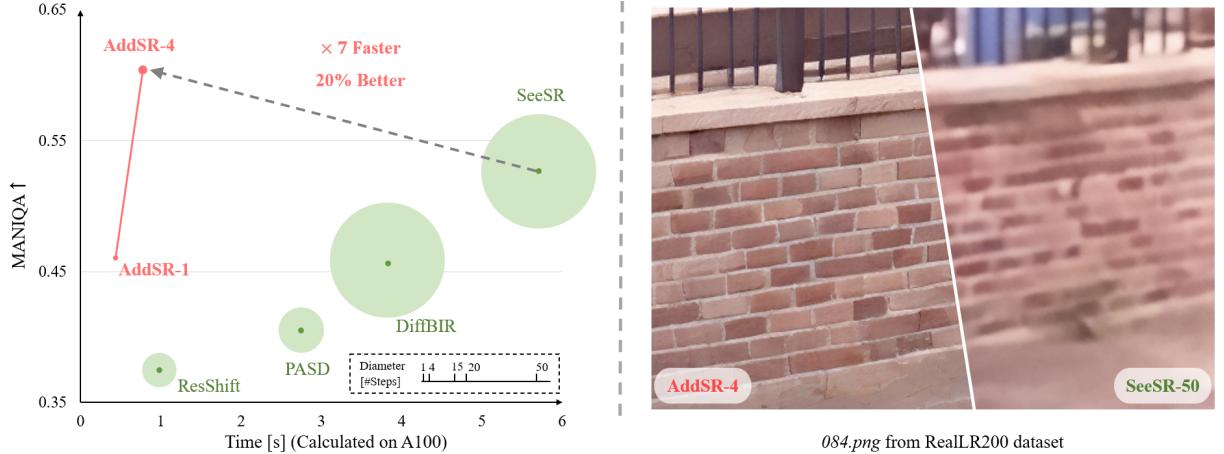


Fig. 1 Comparisons on effect and efficiency. AddSR-4 indicates the result is obtained in 4 steps, achieving high perception quality restoration performance with the fastest speed among diffusion-based models. In contrast, existing SD-based BSR models suffer from either low perception quality restoration performance (*e.g.*, ResShift) or time-consuming efficiency (*e.g.*, SeeSR-50).

input LR images to that of HR images with adversarial training. However, when handling natural images with intricate textures, they often struggle to generate unsatisfactory visual results due to unstable adversarial objectives Liang et al (2022); Xie et al (2023).

Recently, diffusion models (DM) Ho et al (2020); Song et al (2020); Nan et al (2024) have garnered significant attention owing to their potent generative capabilities and the ability to combine information from multiple modalities. DM-based BSR methods can be roughly divided into two categories: those without Stable Diffusion (SD) prior Yue et al (2023); Saharia et al (2023); Li et al (2022), and those incorporating SD prior Lin et al (2023); Wu et al (2023); Wang et al (2023b). SD prior can significantly enhance the model’s ability to capture the distribution of natural images Yang et al (2023), thereby enabling the generated HR images with realistic details. Given the iterative refinement nature of DM, diffusion-based methods typically outperform GAN-based ones, albeit at the expense of efficiency. Hence, there’s an urgent demand for BSR models that *deliver exceptional restoration quality while maintaining high efficiency* for real-world applications.

To achieve the above goal, we draw inspiration from Adversarial Diffusion Distillation (ADD) Sauer et al (2023b) and introduce it into the BSR task. However, two key challenges still

exist: 1) *Perception-distortion imbalance* Blau and Michaeli (2018); Zhang et al (2022b); Luo et al (2024b): Directly applying ADD in the BSR task leads to reduced fidelity, causing a perception-distortion imbalance that undermines effectiveness. 2) *Efficient restoration of high-frequency details*: The quality of the conditioning input can significantly affect the restored results Liao et al (2024). Previous SD-based methods Wang et al (2023b); Lin et al (2023) rely on additional degradation removal modules to pre-clean LR images for conditioning, which hinders efficiency. Therefore, efficiently obtaining a conditioning input with more high-frequency signals to guide restoration is a key challenge for effective and efficient blind super-resolution (BSR).

In this paper, we propose a novel AddSR based on ADD for blind super-restoration, which enhances restoration effects and accelerates inference speed of SD-based models simultaneously. There are two critical designs in AddSR to address the above issues respectively: 1) We introduce *timestep-adaptive adversarial diffusion distillation (TA-ADD)* loss, which designs a bivariate timestep-related weighting function to achieve perception-distortion balance, enhancing generative ability at smaller inference steps while reducing it at larger ones. 2) We propose a simple yet effective strategy, *prediction-based self-refinement (PSR)*, which uses the estimated HR image from the predicted noise to control the model output.

This approach enables efficient condition restoration of the high-frequency components and further allows the restored results to contain more high-frequency details. Our main contributions can be summarized as threefold:

- To the best of our knowledge, the proposed AddSR is the first to explore ADD for efficient and effective blind super-resolution, achieving a $\times 7$ speedup over SeeSR [Wu et al \(2023\)](#) while delivering improved perceptual quality.
- We introduce a new TA-ADD loss to address the perception-distortion imbalance issue introduced by the original ADD, allowing AddSR to generate superior perceptual quality while maintaining comparable fidelity.
- We propose a prediction-based self-refinement strategy to efficiently restore condition and enable the restored results to generate more details without the need for additional modules.

2 Related Work

GAN-based BSR. In recent years, BSR have drawn much attention due to their practicability. Adversarial training [Goodfellow et al \(2014\)](#); [Park et al \(2023\)](#); [Luo et al \(2024a\)](#); [Zhang et al \(2024b\)](#); [Xie et al \(2023\)](#) is introduced in SR task to avoid generating over-smooth results. BSRGAN [Zhang et al \(2021\)](#) designs a random shuffle strategy to enlarge the degradation space for training a comprehensive SR model. Real-ESRGAN [Wang et al \(2021\)](#) presents a more practical degradation process called “high-order” to synthesize realistic LR images. KDSRGAN [Xia et al \(2023b\)](#) estimates the implicit degradation representation to assist the restoration process. While GAN-based BSR methods require only one step to restore the LR image, their capability to super-resolve complex natural images is limited. In this work, our AddSR seamlessly attains superior restoration performance based on diffusion model, making it a compelling choice.

Diffusion-based BSR. Diffusion models has demonstrated significant advantages in image generation tasks (*e.g.*, text-to-image). One common approach [Yue et al \(2023\)](#); [Sahak et al \(2023\)](#); [Xia et al \(2023a\)](#); [Wang et al \(2024\)](#) is training a non-multimodal diffusion model from scratch, which takes the concatenation of a LR image and noise as input in every step. Another approach [Yang et al \(2023\)](#); [Lin et al \(2023\)](#); [Wu et al \(2023\)](#); [Wang](#)

[et al \(2023b\)](#); [Sun et al \(2023\)](#) fully leverages the prior knowledge from a pre-trained multimodal diffusion model (*i.e.*, SD model), which requires training a ControlNet and incorporates new adaptive structures (*e.g.*, cross-attention). SD-based methods excel in performance compared to the aforementioned approaches, as they effectively incorporate high-level information. However, the large number of model parameters and the need for numerous sampling steps pose substantial challenges to application in the real world.

Efficient Diffusion Models. Several works [Lu et al \(2022\)](#); [Song et al \(2020\)](#); [Liu et al \(2022\)](#); [Lu et al \(2023\)](#); [Lin et al \(2024\)](#); [Song et al \(2024\)](#) are proposed to accelerate the inference process of DM. Although these methods can reduce the sampling steps from thousands to 20-50, the restoration effect will deteriorate dramatically. Recently, adversarial diffusion distillation [Sauer et al \(2023b\)](#) is proposed to achieve 1~4 steps inference while maintaining satisfactory generating ability. However, ADD was originally designed for the text-to-image task. Considering the multifaceted nature of BSR, such as image quality, degradation, or the trade-off between fidelity and realness, employing ADD to expedite the SD-based model for BSR is non-trivial. In contrast, AddSR introduces two pivotal designs to adapt ADD into BSR tasks, making it both effective and efficient.

3 Methodology

3.1 Overview of AddSR

Network Components. The AddSR training procedure primarily consists of three components: the student model with weights θ , the pretrained teacher model with frozen weights ψ and the discriminator with weights ϕ , as depicted in Fig. 2. Specifically, both the student model and the teacher model share identical structures, with the student model initialized from the teacher model. The student model incorporates a ControlNet [Zhang et al \(2023a\)](#) to receive x_{LR} or predicted \hat{x}_0^{i-1} for controlling the output of the U-Net [Ronneberger et al \(2015\)](#). Furthermore, the student model utilizes RAM [Zhang et al \(2023b\)](#) to obtain representation embeddings c_{rep} , extracting high-level information (*i.e.*, image content) and sends this information to CLIP [Radford et al \(2021\)](#) to generate text embeddings c_{text} . These embeddings help the backbone (U-Net and ControlNet)

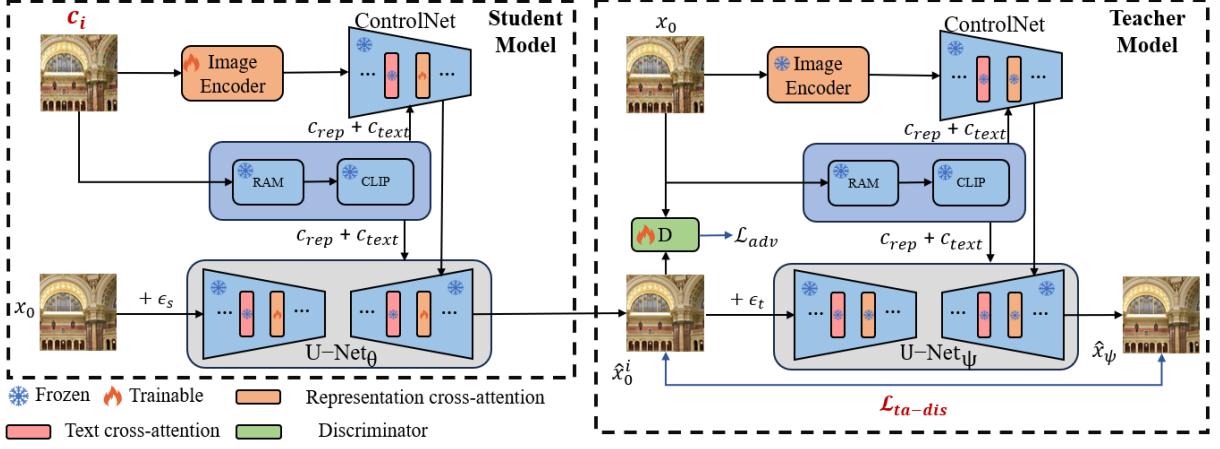


Fig. 2 Overview of AddSR. Our proposed AddSR consists of a student model, a pretrained teacher model, and a discriminator. Let $\mathcal{C} = \{x_{LR}, \hat{x}_0^1, \hat{x}_0^2, \hat{x}_0^3\}$, where c_i denotes the i -th element of \mathcal{C} , \hat{x}_0^i is the predicted HR image in the $(i+1)$ inference step. In our proposed prediction-based self-refinement strategy, we use x_{LR} as the condition only in the first inference step. For the subsequent inference steps, we use the predicted HR image \hat{x}_0^i from the previous step as the condition.

produce high-quality restored images. As for the discriminator, we follow StyleGAN-T Sauer et al (2023a) conditioned on c_{img} extracted from x_{LR} by DINOv2 Oquab et al (2023).

Training Procedure. (1) **Student model with prediction-based self-refinement.** Firstly, we uniformly choose a student timestep s from $\{s_1, s_2, s_3, s_4\}$ (evenly selected from 0 to 999) and employ the forward process on the HR image x_0 to generate the noisy state $x_s = \sqrt{\alpha_s}x_0 + \sqrt{1 - \alpha_s}\epsilon$. Secondly, we input x_s with the condition c_i , the i -th element of $\mathcal{C} = \{x_{LR}, \hat{x}_0^1, \hat{x}_0^2, \hat{x}_0^3\}$ (\hat{x}_0^{i-1} is obtained by PSR to reduce the degradation impact and provide more high-frequency information to the restoration process, as detailed in Sec. 3.2), along with c_{rep} and c_{text} , into the student model to generate samples $\hat{x}_0^i(x_s, s, c_{rep}, c_{text}, c_i)$. (2) **Teacher model.** Firstly, we equally choose a teacher timestep t from $\{t_1, t_2, \dots, t_{1000}\}$ and employ forward process to student-generated samples \hat{x}_θ to obtain the noisy state $\hat{x}_{\theta,t} = \sqrt{\alpha_t}\hat{x}_0 + \sqrt{1 - \alpha_t}\epsilon$. Secondly, we put $\hat{x}_{\theta,t}$ with condition x_0 , c'_{rep} and c'_{text} into teacher model to generate samples $\hat{x}_\psi(\hat{x}_{\theta,t}, t, c'_{rep}, c'_{text}, x_0)$. Note that \hat{x}_ψ is conditioned on x_0 instead of x_{LR} . The primary reason is that substituting x_{LR} with x_0 to regulate the output of the teacher model can force student model implicitly learning the high-frequency information of HR images even conditioned on c_i . (3) **Timestep-adaptive ADD for BSR task.** It consists of two parts: adversarial loss and a novel

timestep-adaptive distillation loss, which is correlated with both the teacher and student model timesteps. The overall objective is:

$$\begin{aligned} \mathcal{L}_{TA-ADD} = & \mathcal{L}_{ta-dis}(\hat{x}_0^i(x_s, s, \rho, c_i), \hat{x}_\psi(\hat{x}_{\theta,t}, t, \rho', x_0), \\ & d(s, t)) + \lambda \mathcal{L}_{adv}(\hat{x}_0^i(x_s, s, \rho, c_i), x_0, \psi_{c_{img}}), \end{aligned} \quad (1)$$

where ρ denotes the c_{rep} and c_{text} , ρ' stands for c'_{rep} and c'_{text} . λ is the balance weight, empirically set to 0.02. $\psi_{c_{img}}$ is the discriminator conditioned on c_{img} . $d(s, t)$ is a weighting function defined by student timestep s and teacher timestep t , dynamically adjusting \mathcal{L}_{ta-dis} and \mathcal{L}_{adv} to alleviate perception-distortion imbalance. Further analysis is provided in Sec. 3.3.

3.2 Prediction-based Self-Refinement

Motivation. As shown in Fig. 3 (a), original SD-based methods directly use LR images to control the output of DM in each inference step. However, some studies Lin et al (2023); Wang et al (2023b); Liao et al (2024) have found that the restored results can be affected by the condition quality, as LR images often suffer from multiple degradations, which can significantly disrupt the restoration process (e.g., see the first line of Fig. 4). To provide a better condition, these methods employ *additional degradation removal models*

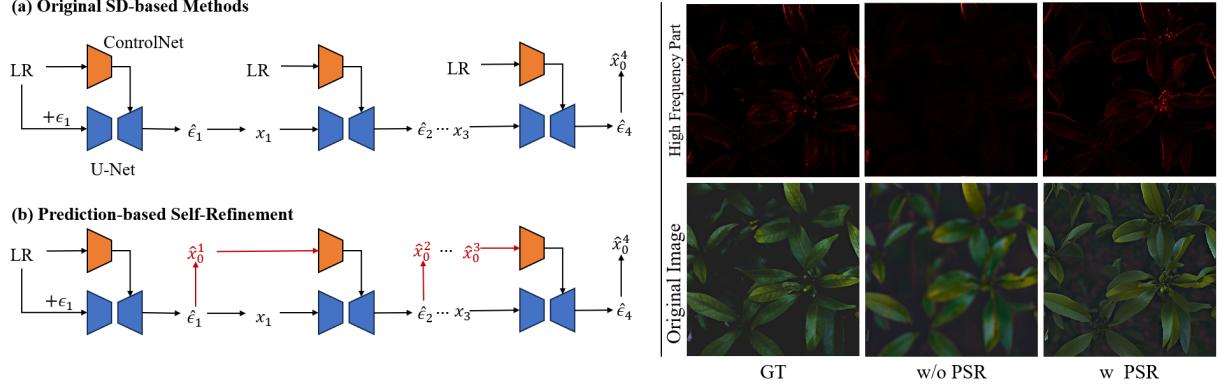


Fig. 3 Illustration of the proposed PSR. The previous SD-based methods usually use LR image to guide model’s output, while our PSR additionally utilizes the predicted HR image from the previous step to provide better supervision with marginal additional time cost.

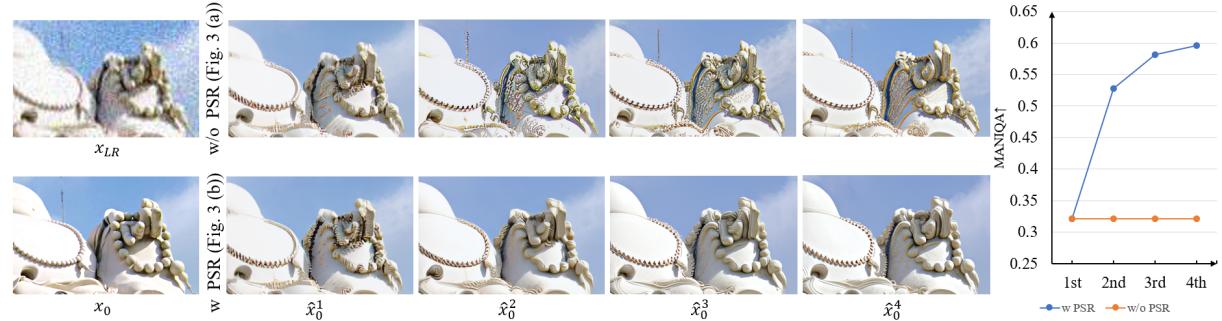


Fig. 4 Left: Visual comparisons with and without PSR. **Right:** Perception quality of the control signal at each timestep. MANIQA is calculated between the input of ControlNet and x_0 .

to pre-clean LR images, aiming to mitigate the impact of degradation. However, such approaches often compromise efficiency, which hinders designing an efficient method.

Approach. To achieve efficient restoration of high-frequency details, we propose a novel prediction-based self-refinement strategy, which incurs only minimal efficiency overhead. The idea of PSR is to utilize the predicted noise to estimate HR. Specifically, we use the following equation:

$$\hat{x}_0 = (x_s - \sqrt{1 - \bar{\alpha}_s} \epsilon_{\theta,s}) / \sqrt{\bar{\alpha}_s} \quad (2)$$

to estimate the HR image \hat{x}_0 from predicted noise in each step, and then control the model output in next step, where x_s is the noisy state and $\epsilon_{\theta,s}$ is the predicted noise at timestep s . The \hat{x}_0 in each step has more high-frequency information to better control the model output (*e.g.*, Fig. 3-right and also Fig. 4-left). Although PSR does not use additional modules to pre-clean LR image, the HR

image estimated by PSR exhibit superior quality compared to the LR image (Fig. 4-right). By leveraging our simple yet effective PSR, AddSR captures conditions with high-frequency information, generating restored results with enhanced detail, without sacrificing efficiency.

3.3 Timestep-Adaptive ADD

Motivation. Perception-distortion trade-off [Blau and Michaeli \(2018\)](#) is a well-known phenomenon in SR task. We observe that training BSR task with ADD directly exacerbates this phenomenon, as shown in Fig. 5(a). Specifically, during the first three inference steps, there is a significant decrease in fidelity, accompanied by improvement in perception quality. In the last inference step, fidelity remains at a low level, while perception quality undergoes a dramatic increase. The aforementioned scenario may give rise to two issues: (1) *When the inference step is small, the quality*

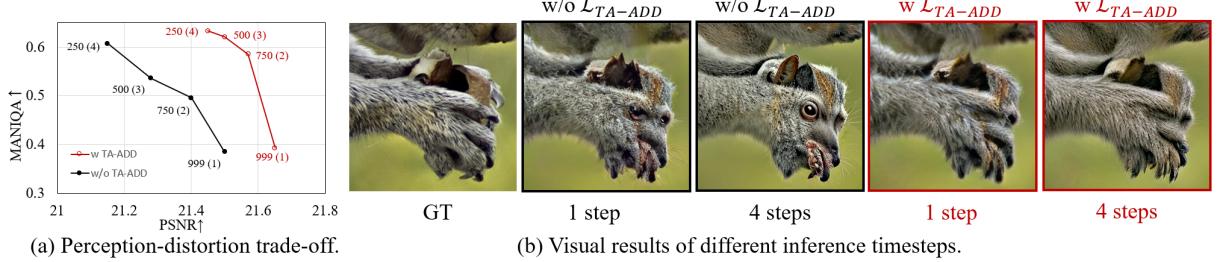


Fig. 5 Illustrations of TA-ADD in perception-distortion trade-off. (a) The perception and fidelity variation trends with and without TA-ADD. MANIQA and PSNR stand for perception quality and fidelity, respectively. (b) The outputs at 1st and 4th timesteps. The final output without \mathcal{L}_{TA-ADD} hallucinates the paw into an animal head, while AddSR retains the appearance of the paw.

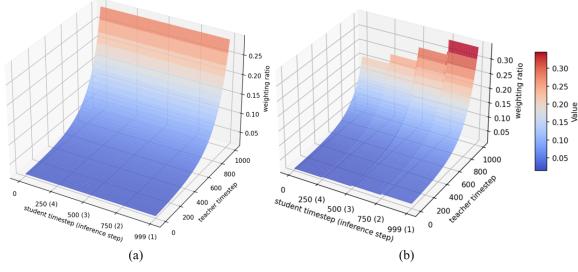


Fig. 6 Relation between weighting ratio and timesteps. (a) weighting ratio $= \lambda / (\prod_{i=0}^t (1 - \beta_t))^{\frac{1}{2}}$. Once the teacher timestep is established, the weighting ratio remains constant. (b) weighting ratio $= \lambda / d(s, t)$. Even the teacher timestep is established, the weighting ratio can change to balance perception-distortion across different student timesteps.

of restored image is subpar. (2) As the inference step increases, the generated images may exhibit “hallucinations”.

Analysis. The primary reason lies with ADD, which maintains a consistent weight for GAN loss and distillation loss across various student timesteps, as depicted in Fig. 6(a). Once the teacher timestep is established, the ratio of adversarial loss and distillation loss remains constant for different student timesteps. However, since the perception quality of generated images gradually increases with larger inference steps, the weight-invariant ADD may lead to insufficient adversarial constraints on the student model during small inference steps, resulting in the generation of blurry images. Conversely, as the inference step increases, the adversarial training constraints become too strong, leading to the generation of “hallucinations” (see Fig. 5(b)).

Approach. To address this issue, we extend the original unary weighting function $(\prod_{i=0}^t (1 - \beta_t))^{\frac{1}{2}}$

to a bivariate weighting function $d(s, t)$, allowing for dynamic adjustment of the ratio between adversarial loss and distillation loss based on both student timestep and teacher timestep, as shown in Fig. 6(b). Specifically, we increase this ratio when only one inference step is performed, and gradually decrease it as the inference step increases. This alleviates the aforementioned issue of generating blurry images with small inference step and “hallucinations” with larger inference steps. We employ the exponential forms to control the weighting ratio. The function $d(s, t)$ is:

$$d(s, t) = \left(\prod_{i=0}^t (1 - \beta_t) \right)^{\frac{1}{2}} \times \mu \cdot \nu^{p(s)-1}, \quad (3)$$

where β represents the noise schedule coefficient, with t and s denoting the teacher timestep and student timestep, respectively. The hyperparameter μ sets the initial weighting ratio, while ν controls the distillation loss increase over student timesteps, typically resulting in higher fidelity with larger ν . The function $p(\cdot)$ serves as a projection function that maps student timesteps to inference steps (e.g., mapping $s = 999$ to 1). We primarily consider the exponential and linear forms to control the weighting ratio. A comparison of detailed settings for different hyper-parameters are provided in Sec. 4.6. From these comparisons, we find that the exponential form of $d(s, t)$ yields good results, so we use Eq. (3) as the distillation loss function for the remaining experiments.

Table 1 Quantitative comparison with SotAs on different degradation cases. ‘*’ indicates that the metric is non-reference. The best results are marked in red, while the second best ones are in blue.

Datasets	Metrics	BSRGAN ICCV 2021	Real-ESRGAN ICCVW 2021	MM-RealSR ECCV 2022	LDL CVPR 2022	FeMaSR MM 2022	StableSR-200 IJCV 2024	ResShift-15 NeurIPS 2023	PASD-20 ECCV 2024	DiffBIR-50 ECCV 2024	SeeSR-50 CVPR 2024	AddSR-1 -	AddSR-4 -
Single: SR($\times 4$)	MANIQA* \uparrow	0.3990	0.3859	0.3959	0.3501	0.4603	0.4088	0.4582	0.4405	0.4680	0.5082	0.3894	0.6430
	MUSIQ* \uparrow	66.06	63.32	64.22	61.10	65.31	65.46	65.50	66.80	67.61	68.88	63.05	71.43
	CLIPQA* \uparrow	0.5951	0.5367	0.5967	0.5120	0.6773	0.6483	0.6803	0.6396	0.6934	0.7039	0.5572	0.7794
	NIQE* \downarrow	5.01	5.21	5.22	5.39	5.80	5.35	5.74	4.68	4.88	5.06	5.31	4.75
	LPIPS \downarrow	0.2003	0.1962	0.1934	0.1892	0.1770	0.1944	0.1544	0.1891	0.2388	0.3085	0.2872	0.2812
	PSNR \uparrow	25.52	25.30	24.35	25.09	23.74	24.45	25.53	25.15	23.43	24.61	22.70	21.83
	SSIM \uparrow	0.7091	0.7158	0.7232	0.7282	0.6788	0.6904	0.7206	0.6896	0.6025	0.6709	0.6012	0.5651
Mixture: Blur($\sigma=2$)+ SR($\times 4$)	MANIQA* \uparrow	0.3823	0.3688	0.3796	0.3337	0.4184	0.3587	0.4195	0.4124	0.4648	0.4974	0.3779	0.6340
	MUSIQ* \uparrow	64.73	60.89	62.21	58.64	62.96	60.85	62.02	64.41	67.09	68.27	61.95	71.11
	CLIPQA* \uparrow	0.5752	0.5116	0.5687	0.4910	0.6390	0.5819	0.6375	0.6026	0.6857	0.6892	0.5389	0.7727
	NIQE* \downarrow	5.17	5.54	5.66	5.72	5.62	5.96	6.25	5.01	5.18	5.32	5.81	6.11
	LPIPS \downarrow	0.2240	0.2267	0.2295	0.2226	0.1979	0.2384	0.2029	0.2223	0.2522	0.2124	0.3007	0.2953
	PSNR \uparrow	25.07	24.74	24.20	24.45	24.00	24.01	24.95	24.70	22.97	24.12	22.57	21.69
	SSIM \uparrow	0.6820	0.6890	0.6973	0.6730	0.6596	0.6926	0.6688	0.5802	0.5905	0.5905	0.5556	
Mixture: SR($\times 4$)+ Noise($\sigma=40$)	MANIQA* \uparrow	0.2645	0.3120	0.3285	0.3138	0.3123	0.3485	0.3741	0.4270	0.4121	0.5537	0.4320	0.6517
	MUSIQ* \uparrow	50.47	53.43	56.53	53.30	56.55	52.24	60.99	64.20	61.85	70.32	65.54	71.26
	CLIPQA* \uparrow	0.4543	0.4761	0.5158	0.6208	0.5178	0.4414	0.5949	0.5503	0.6149	0.7557	0.6219	0.7768
	NIQE* \downarrow	7.04	6.00	4.40	5.61	4.27	5.12	6.10	5.02	5.09	4.95	4.87	6.29
	LPIPS \downarrow	0.4611	0.3601	0.3052	0.3138	0.3267	0.4017	0.3129	0.3451	0.3404	0.2999	0.3546	0.3488
	PSNR \uparrow	17.90	21.97	22.04	22.68	21.84	21.20	22.78	22.12	22.22	21.04	21.01	20.79
	SSIM \uparrow	0.5210	0.6044	0.5998	0.5838	0.5421	0.5077	0.5979	0.5587	0.5311	0.5388	0.5684	0.5621
Mixture: Blur($\sigma=2$)+ SR($\times 4$)+ Noise($\sigma=20$)+ JPEG(q=50)	MANIQA* \uparrow	0.3524	0.3374	0.3287	0.3082	0.3271	0.3452	0.3702	0.4024	0.4538	0.5266	0.3930	0.6335
	MUSIQ* \uparrow	59.83	55.54	55.30	52.79	60.87	61.21	56.99	63.25	64.50	69.08	62.69	70.59
	CLIPQA* \uparrow	0.5380	0.5047	0.4978	0.4699	0.6061	0.6010	0.5888	0.5733	0.6626	0.7180	0.5669	0.7703
	NIQE* \downarrow	5.31	5.69	5.70	5.77	4.87	6.33	7.03	5.49	4.93	5.06	5.13	4.68
	LPIPS \downarrow	0.3223	0.3346	0.3372	0.3272	0.2922	0.3429	0.3526	0.3482	0.3502	0.3085	0.3398	0.3368
	PSNR \uparrow	23.04	22.70	22.47	22.36	22.17	22.39	22.36	22.25	21.46	21.86	21.65	21.45
	SSIM \uparrow	0.5866	0.5935	0.5950	0.5948	0.5633	0.5704	0.5574	0.5594	0.5029	0.5474	0.5312	0.5210

4 Experiments

4.1 Experimental Settings

Training Datasets. We adopt DIV2K [Agustsson and Timofte \(2017\)](#), Flickr2K [Timofte et al \(2017\)](#), first 20K images from LSDIR [Li et al \(2023\)](#) and first 10K face images from FFHQ [Karras et al \(2019\)](#) for training. We use the same degradation model as Real-ESRGAN [Wang et al \(2021\)](#) to synthesize HR-LR pairs.

Test Datasets. We evaluate AddSR on 4 datasets: DIV2K-val [Agustsson and Timofte \(2017\)](#), DRealSR [Wei et al \(2020\)](#), RealSR [Cai et al \(2019\)](#) and RealLR200 [Wu et al \(2023\)](#). We conduct 4 degradation types on DIV2K-val to comprehensively assess AddSR, and except RealLR200, all datasets are cropped to 512×512 and degraded to 128×128 LR image.

Implementation Details. We adopt SeeSR [Wu et al \(2023\)](#) as the teacher model. Note that our approach is applicable to most of the existing SD-based BSR methods for improving restoration results and acceleration. The student model is initialized from the teacher model, and finetuned with Adam optimizer for 50K iterations. The batch size and learning rate are set to 6 and 2×10^{-5} . AddSR is trained under 512×512 resolution with 4 NVIDIA A100 GPUs (40G).

Evaluation Metrics. We employ non-reference metrics (*i.e.*, MANIQA [Yang et al \(2022\)](#), MUSIQ [Ke et al \(2021\)](#), CLIPQA [Wang et al \(2023a\)](#)) and reference metrics (*i.e.*, LPIPS [Zhang](#)

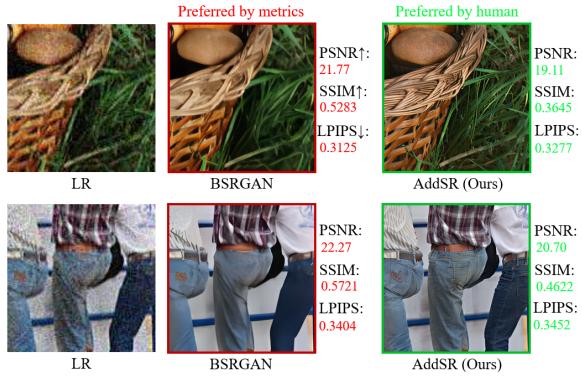


Fig. 7 Illustration on disparity between full-reference metrics and human preference. Despite AddSR achieves lower scores in full-reference metrics, it generates human-preferred images.

[et al \(2018\)](#), PSNR, SSIM [Wang et al \(2004\)](#)) to comprehensively evaluate AddSR. Non-reference metrics are prioritized as they closely align with human perception.

Compared Methods. Extensive state-of-the-art BSR methods are compared, including GAN-based methods: BSRGAN [Zhang et al \(2021\)](#), Real-ESRGAN [Wang et al \(2021\)](#), MM-RealSR [Mou et al \(2022\)](#), LDL [Liang et al \(2022\)](#), FeMaSR [Chen et al \(2022\)](#) and diffusion-based methods: StableSR [Wang et al \(2023b\)](#), ResShift [Yue et al \(2023\)](#), PASD [Yang et al \(2023\)](#), DiffBIR [Lin et al \(2023\)](#), SeeSR [Wu et al \(2023\)](#).

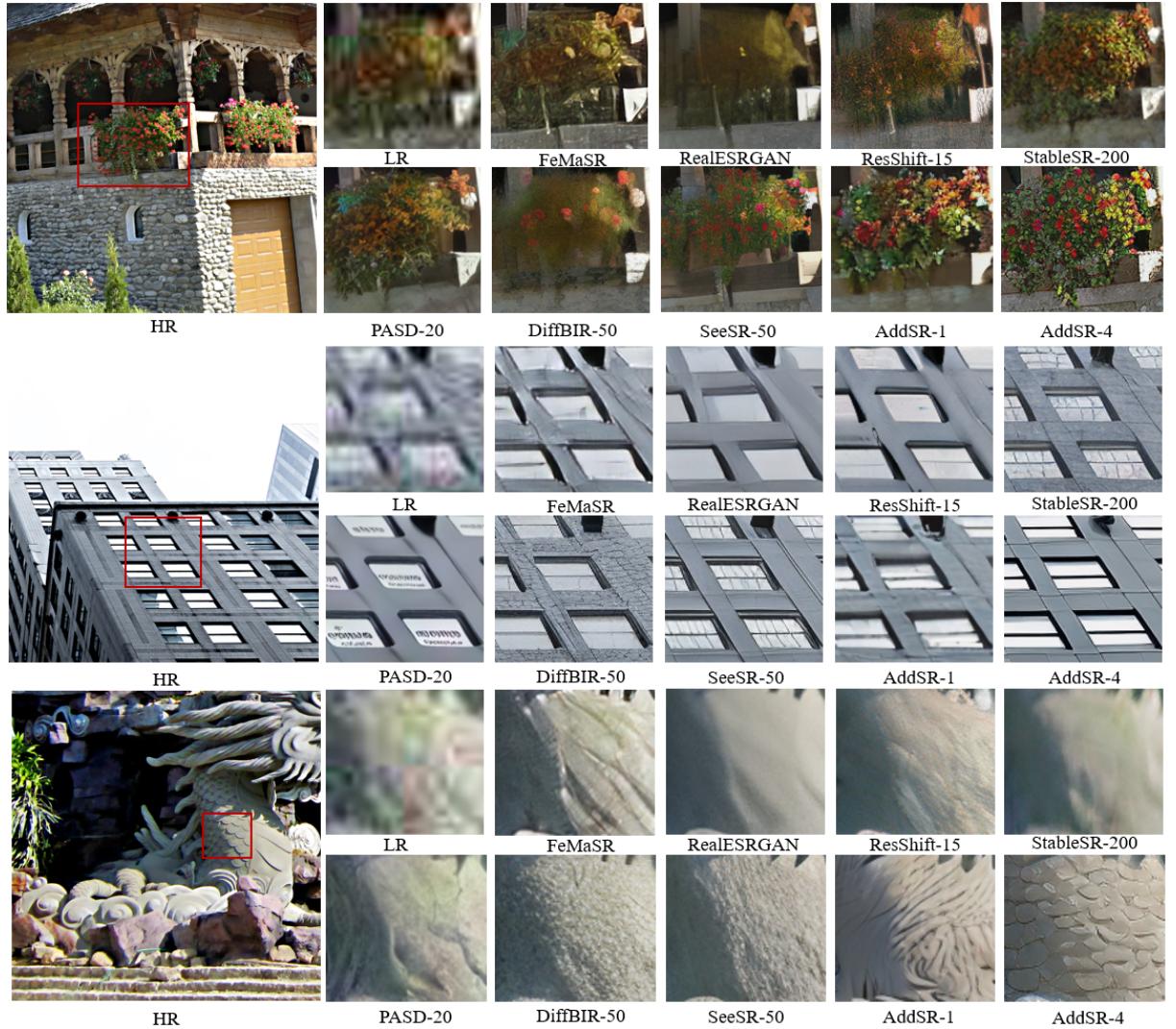


Fig. 8 Visual comparisons on synthetic LR images. Please zoom in for a better view.

Table 2 Quantitative comparison of SUPIR: Inference time, model size, training source, and metrics.

Model	#Params [B]	Time [s]	Training Source	Dataset Size [M]	PSNR↑	SSIM↑	MANIQA↑	CLIPQA↑
SUPIR (CVPR 2024)	~15.56	14.17	64 A6000 (48G)	20	20.78	0.4587	0.6787	0.7992
AddSR (Ours)	~2.28	0.80	4 A100 (40G)	0.034	21.45	0.5210	0.6335	0.7703

4.2 Evaluation on Synthetic Data

To demonstrate the superiority of the proposed AddSR in handling various degradation cases, we synthesized 4 test datasets using the DIV2K-val dataset with different degradation processes. The quantitative results are summarized in Tab. 1. Since SD-based methods emphasize perceptual quality, we provide results using perceptual-priority parameters. In the ablation

study (Tab. 10), we present the corresponding results under balanced parameters. The conclusions include: (1) Our AddSR-4 achieves the highest scores in MANIQA, MUSIQ and CLIP-IQA across 4 degradation cases. Especially for MANIQA, AddSR surpasses the second-best method by more than 16% on average. (2) Diffusion-based models usually achieve low scores in full-reference metrics like PSNR, SSIM and

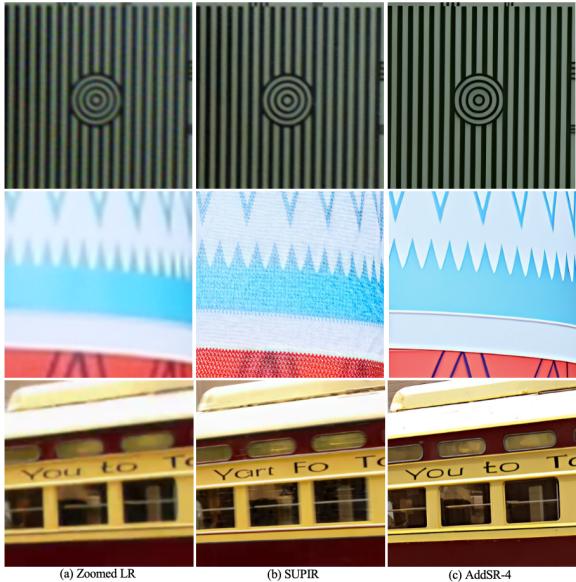


Fig. 9 Visual comparisons with SUPIR. Please zoom in for a better view.

LPIPS, possibly because of their powerful generative ability for realistic details that do not exist in GT. However, full-reference metrics cannot precisely reflect human preferences (see Fig. 7), as discussed in Yu et al (2024); Jinjin et al (2020); Gu et al (2022). (3) AddSR-1 can generate comparative results against other SD-based methods except SeeSR, but significantly reduces the sampling steps (*i.e.*, from ≥ 15 steps to only 1 step).

For a more intuitive comparison, we provide visual results in Fig. 8. One can see that GAN-based method like FeMaSR fails to reconstruct the clean and detailed HR images of the three displayed LR images. As for SD-based method DiffBIR, it tends to generate wrong texture. This is mainly because DiffBIR uses a degradation removal structure to remove the degradation of LR images. However, the processed LR image is blurry, which may lead to the blurry results. Thanks to our proposed PSR, AddSR uses the predicted \hat{x}_0^{i-1} to control the model output, which has more high-frequency information and nearly no extra time cost. With TA-ADD, AddSR can generate precise images and rich details. In a nutshell, AddSR produces images with better perceptual quality than the state-of-the-art models while requiring fewer inference steps and less time.

Moreover, we provide the comparison with SOTA perceptual method SUPIR Yu et al (2024)

in Tab. 2, which details parameters, inference time, training sources, training data, and metrics. As shown, SUPIR exhibits better perceptual quality compared to AddSR. However, AddSR strikes a better balance among model size, inference time, fidelity, perceptual quality and training resource consumption. Although SUPIR demonstrates powerful restoration capabilities, it may still fail in certain degradation scenarios, as shown in the first example of Fig. 9. Furthermore, due to its strong generative ability, the restored results may include details that do not align with the original LR images, such as the texture in the second example and the English letters in the third example in Fig. 9.

4.3 Evaluation on Real-World Data

Tab. 3 shows the quantitative results on 3 real-world datasets. We can see that our AddSR achieves the best scores in MANIQA, MUSIQ and CLIPICQA, the same as in the synthetic degradation cases. This demonstrates that AddSR has an excellent generalization ability to handle unknown complex degradations, making it practical in real-world scenarios. Additionally, AddSR-1 surpasses the GAN-based methods, primarily due to the integration of diffusion model with adversarial training. This integration enables AddSR to leverage the powerful priors of the pre-trained diffusion model and inject high-level information, enhancing the restoration process and producing high-quality perceptual images, even in a *one-step* inference.

Fig. 10 and Fig. 11 show the visualization results. We present the examples of building and face to comprehensively compare various methods. A noticeable observation is that AddSR generates more clear and regular line, as evidenced by the linear pattern of the building in the first example. In the second example, the original LR image is heavily degraded, FeMaSR and ResShift fail to generate the human face, showing only the blurry outline of the face. DiffBIR can generate more details, yet still unclear. The image generated by SeeSR exhibits artifacts. Conversely, our AddSR can generate comparative results with FeMaSR and ResShift in one-step. As evaluating the inference steps, AddSR generates more clear and detailed human face, which significantly surpasses the aforementioned methods.

Table 3 Quantitative comparison with state of the arts on real-world LR images.

Datasets	Metrics	BSRGAN ICCV 2021	Real-ESRGAN ICCVW 2021	MM-RealSR ECCV 2022	LDL CVPR 2022	FeMaSR	StableSR-200 IJCV 2024	ResShift-15 NeurIPS 2023	PASD-20 ECCV 2024	DiffBIR-50 ECCV 2024	SeeSR-50 CVPR 2024	AddSR-1	AddSR-4
RealSR	MANIQA* ↑	0.3762	0.3727	0.3966	0.3417	0.3609	0.3656	0.3750	0.4041	0.4392	0.5396	0.4189	0.6597
	MUSIQ* ↑	63.28	60.36	62.94	58.04	59.06	61.11	56.06	62.92	64.04	69.82	63.56	72.25
	CLIP1QA* ↑	0.5116	0.4492	0.5281	0.4295	0.5408	0.5277	0.5421	0.5187	0.6491	0.6700	0.4929	0.7215
	LPIPS↓	0.2670	0.2727	0.2783	0.2766	0.2942	0.3018	0.3460	0.3435	0.3658	0.3007	0.3095	0.3742
DrealSR	PSNR↑	26.49	25.78	23.69	25.09	25.17	25.63	26.34	26.67	25.06	25.24	24.22	22.73
	SSIM↑	0.7667	0.7621	0.7470	0.7642	0.7359	0.7483	0.7352	0.7577	0.6664	0.7204	0.6863	0.6336
	MANIQA* ↑	0.3431	0.3428	0.3625	0.3237	0.3178	0.3222	0.3284	0.3874	0.4646	0.5125	0.3873	0.6034
	MUSIQ* ↑	57.17	54.27	56.71	52.38	53.70	52.28	50.14	55.33	60.40	65.08	57.42	68.16
RealLR200	CLIP1QA* ↑	0.5094	0.4514	0.5171	0.4410	0.5639	0.5101	0.5287	0.5384	0.6397	0.6910	0.5543	0.7381
	LPIPS↓	0.2883	0.2847	0.2832	0.2815	0.3169	0.3315	0.4006	0.3931	0.4599	0.3299	0.3025	0.3866
	PSNR↑	28.68	28.57	26.84	27.41	26.83	29.14	28.27	29.06	26.56	28.09	27.49	26.09
	SSIM↑	0.8022	0.8042	0.7959	0.8069	0.7545	0.8040	0.7542	0.7906	0.6436	0.7664	0.7588	0.7036
MANIQA* ↑		0.3688	0.3656	0.3879	0.3266	0.4099	0.3672	0.4182	0.4193	0.4626	0.4911	0.4215	0.6182
MUSIQ* ↑		64.87	62.93	65.24	60.95	64.24	62.89	60.25	66.35	66.84	68.63	65.02	72.62
CLIP1QA* ↑		0.5699	0.5423	0.6010	0.5088	0.6547	0.5916	0.6468	0.6203	0.6965	0.6617	0.5679	0.7724

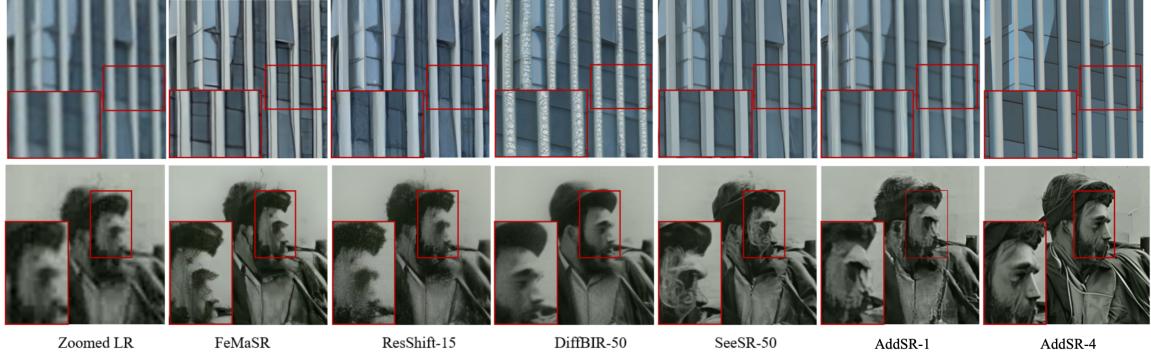


Fig. 10 Visual comparisons on real-world LR images from RealLR200. Please zoom in for a better view.



Fig. 11 Visual comparisons on real-world LR images from RealSR and DrealSR. Zoom in for a better view.

Prompt-Guided Restoration. One of the advantages of diffusion model is to integrate with text. In Fig. 12, we demonstrate that our AddSR can efficiently achieve more precise restoration results *in 4 steps* by incorporating with manual prompt, *i.e.*, we can manually input

the text description of the LR image to assist the restoration process. Specifically, Fig. 12(a) demonstrates that the plaid shirt in the restored image using the RAM prompt can be edited to camouflage with a manual prompt. In Fig. 12(b), the Spider-Man restored with the original RAM

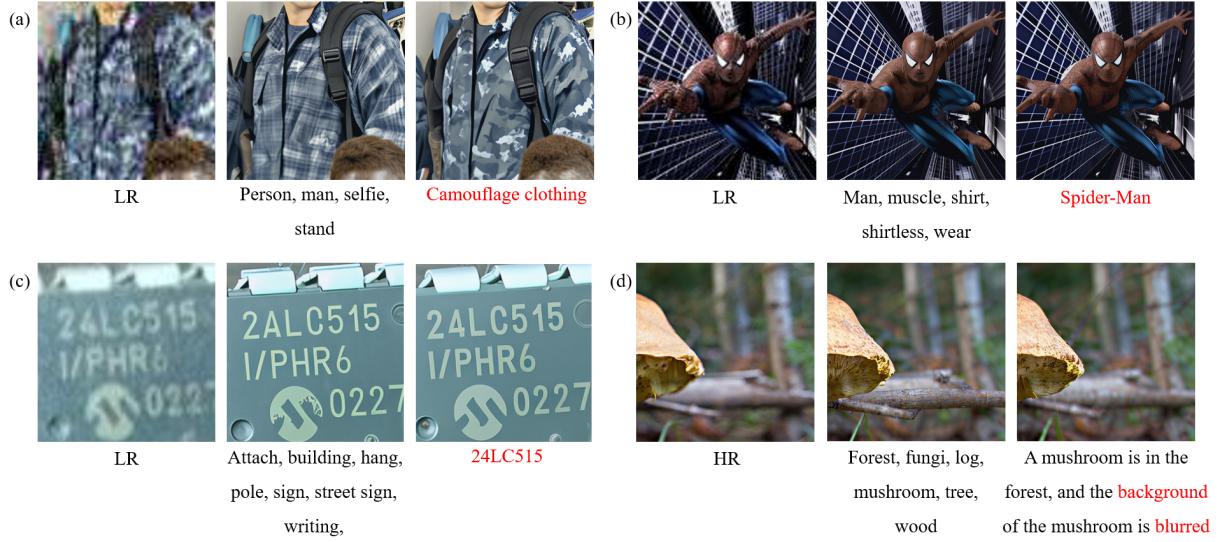


Fig. 12 Illustrations of prompt-guided restoration that engages with manual prompts for more precise outcomes. In each group, the prompts for the second and third images are obtained through RAM and manual input, respectively.

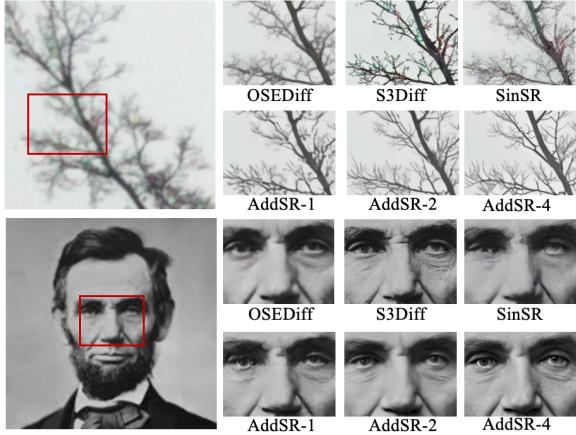


Fig. 13 Qualitative comparison with other efficient methods. AddSR achieves comparable results within a single step and surpasses other efficient methods after 2 or 4 steps (Zoom in for details).

prompt features a mouth and beard, whereas the manual prompt accurately restores the man wearing a spiderweb suit. Fig. 12(c) shows that the text on the chip can be corrected from ‘2ALC515’ to ‘24LC515’ using the manual prompt. Finally, Fig. 12(d) illustrates that while the RAM prompt renders the tree branches sharply, the manual prompt preserves the intended background blur of the mushroom, aligning with the Ground Truth.

4.4 Comparison with Other Efficient Methods

We conduct a quantitative comparison with several other efficient methods Zhang et al (2024a); Wang et al (2024), along with the Turbo versions of SeeSR Wu et al (2023) and StableSR Wang et al (2023b), as shown in Tab. 4. AddSR-1 achieves comparable performance in MANIQA and MUSIQ metrics on RealLR200, as well as in PSNR, SSIM and MANIQA metrics on DrealSR. The primary reason AddSR-1 does not outperform other efficient methods is that those methods are specifically trained for one-step inference, which gives them an advantage in one-step setting. In contrast, AddSR offers greater flexibility, allowing for the application of multiple inference steps. As shown in Tab. 4, AddSR-2 outperforms other efficient methods in the MANIQA and MUSIQ metrics on RealLR200 and DrealSR datasets. Furthermore, AddSR-4 achieves the best results on both MANIQA and MUSIQ metrics.

We also present a qualitative comparison in Figure 13. The visual results of AddSR-1 are comparable to those of other efficient methods. Thanks to the flexibility of AddSR, we can perform multiple inference steps. AddSR-2 and AddSR-4 produce clearer and more realistic results compared to other efficient methods.

Table 4 Quantitative comparison with other efficient methods, *i.e.*, SeeSR-Turbo Wu et al (2023), StableSR-Turbo Wang et al (2023b), S3Diff Zhang et al (2024a) and SinSR Wang et al (2024). The number following the methods denotes the number of inference steps.

Methods	RealLR200			DrealSR	
	MANIQA↑	MUSIQ↑	PSNR↑	SSIM↑	MANIQA↑
SeeSR-Turbo-2	0.3503	53.88	26.24	<u>0.7531</u>	0.3825
StableSR-Turbo-4	0.4208	67.55	24.70	0.6639	0.3381
S3Diff-1	0.4621	68.92	26.89	0.7469	0.4508
SinSR-1	0.4436	63.76	28.37	0.7486	0.3829
AddSR-1 (Ours)	0.4215	65.02	27.49	0.7588	0.3873
AddSR-2 (Ours)	<u>0.5797</u>	<u>71.44</u>	26.78	0.7232	<u>0.5643</u>
AddSR-4 (Ours)	0.6182	72.62	26.09	0.7036	0.6034

Table 5 Complexity comparison of model complexity. The inference time of each method is calculated on 128×128 input images for 4x SR on an A100 GPU.

	StableSR	DiffBIR	SeeSR	ResShift	SinSR	OSEDiff	S3Diff	AddSR (Ours)
Inference steps	200	50	50	15	1	1	1	1 / 4
Runtime [s]	17.75	16.06	5.64	1.65	0.42	0.60	0.62	0.63 / 0.81

4.5 Complexity Analysis

We compare diffusion-based methods in terms of the number of inference steps and runtime in Tab. 5. The runtime is calculated on an A100 GPU. From this comparison, we can make the following observations: 1) SinSR has the fastest runtime among other one-step methods, as it is not a SD-based method and has fewer total network parameters. 2) Among SD-based methods, OSEDiff, S3Diff and AddSR have similar runtimes in a single inference step, as their network structures are similar.

4.6 Ablation Study

Effectiveness of Refined Training Process. To enrich the information provided by the teacher model, we refine the training process by substituting the LR image with HR image as inputs of ControlNet, RAM and CLIP. Since SeeSR is adopted as the baseline, we also replace the LR image of its RAM input with HR image. The quantitative results are shown in Tab. 6. With the supervision from HR input, the perception quality of restored images becomes better.

Comparison of Timestep-Adaptive ADD Forms. To determine the optimal settings for timestep-adaptive ADD, we conduct experiments on its different forms: exponential and linear. Specifically, the exponential form is defined as

Eq. 3, while the linear form is defined as follows:

$$d(s, t) = \left(\prod_{i=0}^t (1 - \beta_t) \right)^{\frac{1}{2}} \times (\gamma \cdot p(s) + \kappa) \quad (4)$$

where the hyper-parameter κ sets the initial weighting ratio, while γ controls the increase of distillation loss over student timesteps. The quantitative results of the exponential and linear forms under various settings are listed in Tab. 7 and Tab. 8, respectively. The best settings for different forms in the tables are highlighted with a gray background. From these tables, we can draw the following conclusions: (1) The best results of the exponential form are better than those of the linear form. Therefore, we use Eq. (3) as the distillation loss function. Moreover, when $\mu = 0.5$ and $\nu = 2.1$, we achieve the best perceptual quality while maintaining good fidelity, so we use this setting for Eq. (3). (2) Increasing the hyper-parameters that control the distillation loss ratio (*i.e.*, ν and γ) typically results in higher fidelity. For instance, when we fix μ to 0.5 and increase ν , the overall trend in 4 step shows a decrease in perception quality and an improvement in fidelity. Consequently, we can achieve a perception-distortion trade-off by adjusting ν .

Effectiveness of TA-ADD on Balancing Perception and Fidelity. The proposed TA-ADD aims to balance perception and fidelity quality of restored images. The quantitative results are shown in Tab. 9. Despite we increase the weight

Table 6 Ablation studies on refined training process. The best results are marked in **bold**.

Exp	Condi Image	RAM	RealLR200			DrealSR			
			MANIQA*↑	MUSIQ*↑	CLIPQA*↑	MANIQA*↑	MUSIQ*↑	CLIPQA*↑	PSNR↑
(1)	✗	✗	0.5623	70.75	0.7431	0.5331	64.55	0.7285	26.96
(2)	✓	✗	0.6092	71.76	0.7660	0.5372	62.67	0.6997	26.78
(3)	✗	✓	0.5772	71.33	0.7549	0.5433	65.09	0.7087	26.87
AddSR	✓	✓	0.6182	72.62	0.7724	0.6034	68.16	0.7381	26.09

Table 7 Comparing the exponential form of timestep-adaptive ADD across different hyper-parameters.

μ	ν	RealSR							
		1 step		2 step		3 step		4 step	
		MANIQA↑	PSNR↑	MANIQA↑	PSNR↑	MANIQA↑	PSNR↑	MANIQA↑	PSNR↑
0.5	1.3	0.4202	24.11	0.6263	23.10	0.6427	22.36	0.6453	22.33
	1.7	0.3986	24.18	0.6110	24.01	0.6195	23.44	0.6307	23.40
	2.1	0.4189	24.22	0.6339	23.29	0.6496	22.76	0.6597	22.73
	2.5	0.4207	24.48	0.5939	23.92	0.6081	23.33	0.6197	23.29
0.7	2.1	0.3821	24.90	0.5971	24.03	0.6078	23.39	0.6221	23.36
0.9		0.4095	23.16	0.6052	22.97	0.6062	22.88	0.6244	22.68

of the distillation loss in the later inference steps, the perceptual quality still improves. This could be attributed to the initial steps producing sufficiently high perception quality images, which offer more informative cues when combined with PSR. Consequently, the later inference steps can achieve high perceptual quality.

In addition, we can adjust the hyperparameters in TA-ADD and the number of inference steps to achieve competitive PSNR and SSIM results while excelling in perceptual quality. We primarily compare the leading methods in terms of fidelity (GAN-based Real-ESRGAN) and perceptual quality (SeeSR). As shown in Tab. 10, Our method remarkably enhances perceptual quality compared to SeeSR while also offering better fidelity. When compared to Real-ESRGAN, our method shows a substantial improvement in perceptual quality while maintaining comparable fidelity. This indicates that TA-ADD effectively navigates the perception-fidelity trade-off. Specifically, we made the following adjustments: 1) larger values for μ and ν ($\mu=0.7, \nu=2.1$) in TA-ADD during training, and 2) fewer inference steps (2 steps) to achieve high-fidelity results.

The visual results are shown in Fig. 14-Right. For the upper 3 images, the content is a statue. However, without TA-ADD, the model hallucinates its hand as a bird. For the bottom 3 images, the original background is rock. Again, without utilizing TA-ADD, AddSR might hallucinate

the background as an eye of a wolf. Conversely, with the help of TA-ADD, the restored images can generate more consistent contents with GTs. TA-ADD constrains the model from excessively leveraging its generative capabilities, thereby preserving more information in the image content, aligning closely with the GTs. Specifically, using TA-ADD, texture of the statue’s hand in the upper image remains unchanged, and the background of the bottom image retains the rock with out-of-focus appearance.

Effectiveness of PSR. As shown in Tab. 9, incorporating PSR significantly enhances perceptual quality with minimal computational cost. All of the three perception metrics, including MANIQA, MUSIQ and CLIPQA, are improved on the two popular real-world datasets, namely RealLR200 and DrealSR. Moreover, we use a additional degradation removal model to replace the PSR to demonstrate the effectiveness of PSR. Following StableSR Wang et al (2023b) and DiffBIR Lin et al (2023), we use Real-ESRGAN Wang et al (2021) or SwinIR Liang et al (2021) to replace the PSR. The quantitative results are shown in Table 11. We find that using Real-ESRGAN or SwinIR achieve comparable results to those obtained using our proposed PSR. However, the use of Real-ESRGAN or SwinIR compromises efficiency, as it requires an additional degradation removal model.

We also attempt to apply PSR to other SD-based methods without fine-tuning. However, we

Table 8 Comparing the linear form of timestep-adaptive ADD across different hyper-parameters.

γ	κ	RealSR							
		1 step		2 step		3 step		4 step	
		MANIQA↑	PSNR↑	MANIQA↑	PSNR↑	MANIQA↑	PSNR↑	MANIQA↑	PSNR↑
0.1	0.7	0.3908	24.28	0.5849	23.64	0.6133	23.00	0.6172	22.98
	0.3	0.4574	23.69	0.6294	23.06	0.6465	22.48	0.6480	22.41
	0.2	0.4338	23.87	0.6237	23.17	0.6407	22.55	0.6443	22.52
0.7	0.7	0.4225	24.02	0.6064	23.24	0.6313	22.60	0.6352	22.59
	0.1	0.4027	24.41	0.6077	23.25	0.6157	22.51	0.6215	22.45
	0.4	0.4045	24.39	0.5981	23.30	0.6124	22.58	0.6202	22.51
0.6	0.5	0.4152	24.73	0.6261	23.33	0.6495	22.65	0.6507	22.62
	0.1	0.4156	24.48	0.6175	23.44	0.6261	22.82	0.6328	22.79
	0.3	0.3926	24.90	0.5905	23.93	0.5857	23.51	0.5981	23.42
0.8	0.1	0.3887	24.92	0.5943	23.84	0.6038	23.36	0.6130	23.28

Table 9 Ablation studies on PSR and TA-ADD.

Methods	Time[s]	RealLR200			DrealSR			
		MANIQA* ↑	MUSIQ* ↑	CLIPQA* ↑	MANIQA* ↑	MUSIQ* ↑	CLIPQA* ↑	PSNR↑
w/o PSR	0.63~0.77	0.5863	72.07	0.7631	0.5752	66.67	0.7243	26.52
w/o TA-ADD	0.63~0.81	0.6058	72.19	0.7630	0.5898	67.58	0.7042	25.77
AddSR	0.63~0.81	0.6182	72.62	0.7724	0.6034	68.16	0.7381	26.09

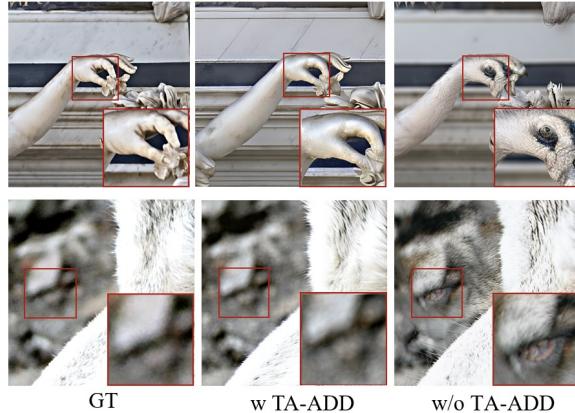


Fig. 14 Visual comparison of TA-ADD.

find that directly applying PSR to existing SD-based methods often results in collapsed outputs. The main reasons for this phenomenon can be summarized as follows: 1) Other SD-based methods typically require dozens of steps, which causes the initially predicted HR image to deviate significantly from the ground truth (GT). As the denoising process progresses, these errors accumulate, ultimately leading to significant deviations in the final results. 2) Due to the lack of fine-tuning, the distribution of the ControlNet input predicted HR image differs significantly from the

LR images used during pre-training. This discrepancy prevents the model from effectively handling this condition, ultimately leading to the collapse of the final outputs.

5 Inference Strategies

In this section, we discuss the impact of different inference strategies on AddSR. We mainly consider two aspects: 1) the blended PSR, as discussed in Sec. 5.1; and 2) the impact of different positive prompts, as explored in Sec. 5.2.

5.1 Blended Prediction-based Self-Refinement

To mitigate error accumulation caused by the predicted HR image, we can use a mixture of the LR image and the predicted HR image to control the model output, rather than relying solely on the predicted HR image. This approach can be formulated as follows:

$$x_{mix} = r\hat{x}_0^i + (1-r)x_{LR} \quad (5)$$

where x_{mix} denotes mixture of \hat{x}_0 and x_{LR} , r is blending ratio. As shown in Tab. 12, a ratio of 0 indicates that only the LR image is used

Table 10 Quantitative comparison of different settings for TA-ADD on synthetic degraded DIV2K. The best results are **bold**, and the second best results are underlined.

Metrics	Real-ESRGAN	SeeSR	Ours-perception (default)	Ours-fidelity
MANIQA↑	0.3374	0.5266	0.6335	<u>0.5759</u>
CLIPIQ↑	0.5047	0.7180	0.7703	<u>0.7357</u>
PSNR↑	22.70	21.86	21.45	<u>21.92</u>
SSIM↑	0.5935	0.5474	0.5210	<u>0.5481</u>

Table 11 Comparison on different degradation removal methods.

Methods	RealLR200			DrealSR		
	MANIQA↑	MUSIQ↑	CLIPIQ↑	PSNR↑	SSIM↑	MANIQA↑
w Real-ESRGAN	0.5955	72.33	0.7692	26.03	0.6998	0.6012
w SwinIR	0.5956	72.32	0.7679	26.21	0.7112	0.6026
w PSR	0.6182	72.62	0.7724	26.09	0.7036	0.6034

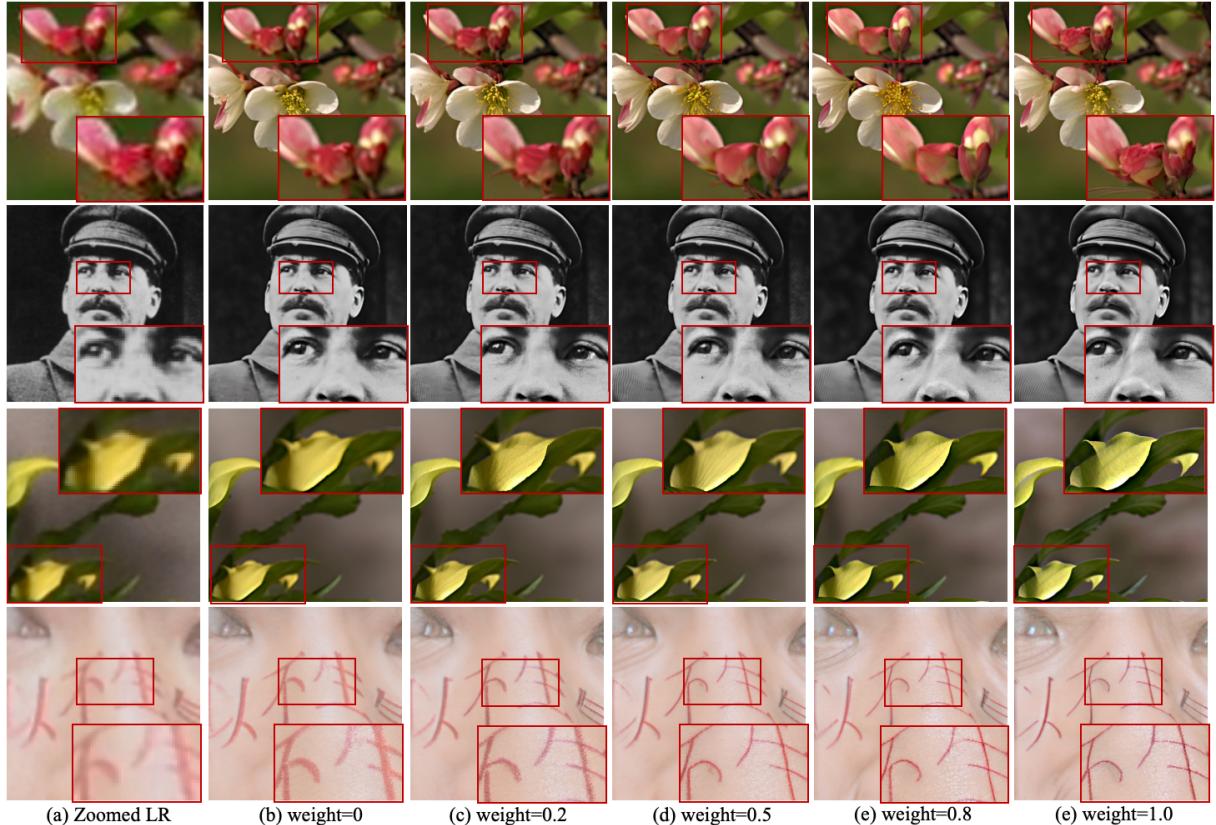


Fig. 15 Qualitative results under different PSR weights on RealLR200 and DrealSR datasets (Zoom in for details).

as the ControlNet input, while a ratio of 1 indicates that only the predicted HR image is used as the ControlNet input. Based on the results in the table, we can draw the following conclusions: 1) Increasing the blending ratio, *i.e.*, increasing the proportion of the predicted HR image,

enhances the perceptual quality. This is evidenced by the gradual improvement of the MANIQA, MUSIQ, and CLIPIQ metrics on both real-world datasets as the ratio increases. This improvement can be attributed to the high-frequency information in the predicted HR image, which enriches the

Table 12 Quantitative results under different blending ratio of PSR. We set the ratio to 1 as the default setting.

Ratio	RealLR200			DrealSR			
	MANIQA↑	MUSIQ↑	CLIPQA↑	PSNR↑	SSIM↑	MANIQA↑	CLIPQA↑
0	0.5863	72.07	0.7631	26.52	0.7342	0.5752	0.7243
0.2	0.5892	72.23	0.7633	26.43	0.7289	0.5801	0.7313
0.5	0.5939	72.37	0.7699	26.27	0.7157	0.5962	0.7324
0.8	0.6016	72.49	0.7707	26.19	0.7099	0.6012	0.7365
1.0	0.6182	72.62	0.7724	26.09	0.7036	0.6034	0.7381

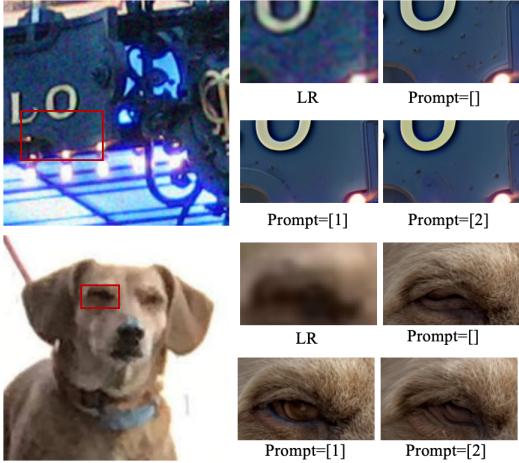


Fig. 16 Qualitative results under different prompts (Zoom in for details).

details of the restored results. 2) Fidelity decreases as the blending ratio increases. Specifically, the PSNR and SSIM metrics on the DrealSR dataset decrease as the blending ratio grows. The primary reason for this is that, although the predicted HR image contains more high-frequency information, it may introduce details that do not exist in the LR image. Such errors accumulate during denoising, ultimately compromising fidelity.

The visual results are shown in Fig. 15. As the blending ratio increases, the results become clearer, as illustrated by the examples in the first and third rows. However, an excessively high ratio may also introduce errors. For instance, in the second and fourth rows, while the results appear clearer and more realistic, certain details deviate from the original LR images, such as the thinner fonts in the fourth-row example.

5.2 Positive Prompts Impact

Since AddSR does not utilize classifier-free guidance Ho and Salimans (2022) during inference, our analysis primarily focuses on the

impact of positive prompts. Following StableSR Wang et al (2023b), we compare three settings: 1) no positive prompt, denoted as $[]$; 2) “clean, high-resolution, 8k, extremely detailed, best quality, sharp”, denoted as [1]; and 3) “Good photo”, denoted as [2]. As shown in Tab. 13, using positive prompt [1] slightly increases the perceptual quality metrics but compromises result fidelity, similar to the effect of using positive prompt [2]. The primary reason is that with a positive prompt, the restored results generate clearer details, improving perceptual quality metrics but deviating from the original LR image, reducing fidelity. As shown in Fig. 16, using positive prompt [1] helps the SR model better remove degradation (e.g., the first example) and generate more accurate structures, such as the dog’s eyes in the second example.

6 Discussion

In this section, we provide a comparison among ADD, SeeSR, and our proposed AddSR. Their architecture diagrams are depicted in Fig. 17. Firstly, the distinctions between ADD and AddSR primarily lie in two aspects: 1) *Introduction of ControlNet*: ADD is originally developed for text-to-image task, which typically only takes text as input. In contrast, AddSR is an image-to-image model that requires the additional ControlNet to receive information from the LR image. 2) *Perception-distortion Trade-off*: ADD aims to generate photo-realistic images from texts. However, introducing ADD into blind SR brings the perception-distortion imbalance issue (please refer to Sec. 3.4 in our submission), which is addressed by our proposed timestep-adaptive ADD in AddSR.

Secondly, the key differences between SeeSR and AddSR are: 1) *Introduction of Distillation*: SeeSR is trained based on vanilla SD model that

Table 13 Comparison of different prompts and guidance strengths. Note that $s = 0$ is equivalent to using negative prompts w/o guidance. Positive prompts are [1] “clean, high-resolution, 8k, extremely detailed, best quality, sharp”, and [2] “Good photo”. The Second row is the default settings for AddSR.

Pos. Prompts	RealLR200			DrealSR			
	MANIQA↑	MUSIQ↑	CLIPQA↑	PSNR↑	SSIM↑	MANIQA↑	CLIPQA↑
[]	0.5895	72.56	0.7674	26.27	0.7139	0.5911	0.7311
[1]	0.6182	72.62	0.7724	26.09	0.7036	0.6034	0.7381
[2]	0.5909	72.56	0.7666	25.95	0.6998	0.5932	0.7345

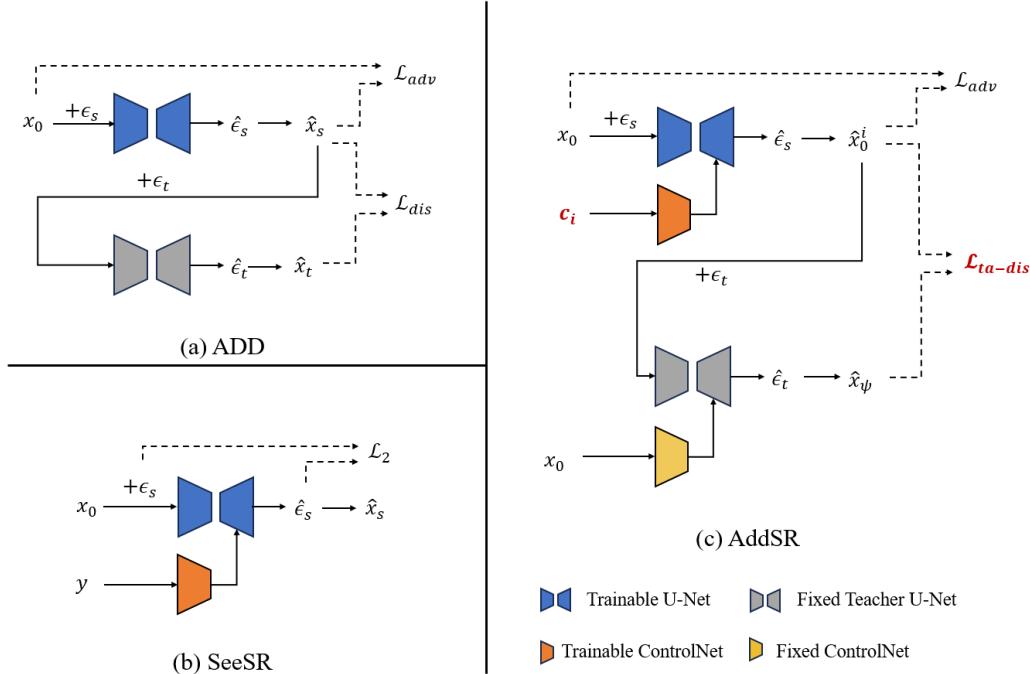


Fig. 17 Comparisons on architecture diagram among ADD, SeeSR and AddSR. x_0 and y denote HR and LR images, respectively. \hat{x}_0^i and \hat{x}_ϕ denote the predicted x_0 from the timesteps s and t , respectively. ϵ_s , ϵ_t , $\hat{\epsilon}_s$ and $\hat{\epsilon}_t$ stand for added and predicted noise in timesteps s and t , respectively. \mathcal{L}_{ta-dis} is the timestep-adaptive distillation loss.

needs 50 inference steps, while AddSR utilizes a teacher model to distill an efficient student model to achieve just 1~4 steps. 2) *High-frequency Information*: SeeSR uses the LR image y as the input of the ControlNet. In contrast, AddSR on one hand adopts the HR image x_0 as the input of the teacher model’s ControlNet to supply the high-frequency signals since the teacher model is not required during inference. On the other hand, AddSR proposes a novel prediction-based self-refinement (PSR) to further provide high-frequency information by replacing the LR image with the predicted image as the input of the student model’s ControlNet. Therefore, AddSR has the ability to generate results with more realistic details.

7 Conclusion

We propose AddSR, an effective and efficient model based on Stable Diffusion prior for blind super-resolution. To address the perception-distortion imbalance issue introduced by the original ADD, we introduce timestep-adaptive ADD, which assigns distinct weights to GAN loss and distillation loss across different student timesteps. In contrast to current SD-based BSR approaches that either use LR images to regulate each inference step’s output or rely on additional modules to pre-clean LR images as conditions, AddSR substitutes the LR image with the HR image estimated in the preceding step. This substitution

provides more high-frequency information, allowing for restored results with enhanced textures and edges, while maintaining efficiency. Additionally, we use the HR image as the controlling signal for the teacher model, enabling it to provide better supervision. Extensive experiments demonstrate that AddSR generates superior results within 1~4 steps in various degradation scenarios and real-world low-quality images.

Limitations. Although the inference speed of our AddSR surpasses all of the existing SD-based methods remarkably, there still exists a gap between AddSR and GAN-based methods. The primary factor is that AddSR is built upon SD and ControlNet, which, due to its substantial model parameters and intricate network structure, noticeably hinders the inference time. In the future, we plan to explore a more streamlined network architecture to boost overall efficiency.

References

- Agustsson E, Timofte R (2017) Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW
- Blau Y, Michaeli T (2018) The perception-distortion tradeoff. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6228–6237, <https://doi.org/10.1109/CVPR.2018.00652>
- Cai J, Zeng H, Yong H, et al (2019) Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE International Conference on Computer Vision
- Chen C, Shi X, Qin Y, et al (2022) Real-world blind super-resolution via feature matching with implicit high-resolution priors
- Chen H, Wang Y, Guo T, et al (2021) Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12299–12310
- Dong C, Loy CC, He K, et al (2016) Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(2):295–307. <https://doi.org/10.1109/TPAMI.2015.2439281>
- Goodfellow IJ, Pouget-Abadie J, Mirza M, et al (2014) Generative adversarial networks. *1406.2661*
- Gu J, Cai H, Dong C, et al (2022) Ntire 2022 challenge on perceptual image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 951–967
- Ho J, Salimans T (2022) Classifier-free diffusion guidance. arXiv preprint arXiv:220712598
- Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33:6840–6851
- Jinjin G, Haoming C, Haoyu C, et al (2020) Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, pp 633–651
- Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 4396–4405, <https://doi.org/10.1109/CVPR.2019.00453>
- Ke J, Wang Q, Wang Y, et al (2021) Musiq: Multi-scale image quality transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 5128–5137, <https://doi.org/10.1109/ICCV48922.2021.00510>
- Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1646–1654, <https://doi.org/10.1109/CVPR.2016.182>
- Ledig C, Theis L, Huszár F, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 105–114, <https://doi.org/10.1109/CVPR.2017.19>

- Li H, Yang Y, Chang M, et al (2022) Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* 479:47–59
- Li Y, Zhang K, Liang J, et al (2023) Lsdir: A large scale dataset for image restoration. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1775–1787, <https://doi.org/10.1109/CVPRW59228.2023.00178>
- Liang J, Cao J, Sun G, et al (2021) Swinir: Image restoration using swin transformer. arXiv preprint arXiv:210810257
- Liang J, Zeng H, Zhang L (2022) Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
- Liao K, Yue Z, Wang Z, et al (2024) Denoising as adaptation: Noise-space domain adaptation for image restoration. URL <https://arxiv.org/abs/2406.18516>, [2406.18516](#)
- Lin S, Wang A, Yang X (2024) Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:240213929
- Lin X, He J, Chen Z, et al (2023) Diffbir: Towards blind image restoration with generative diffusion prior. arxiv
- Liu L, Ren Y, Lin Z, et al (2022) Pseudo numerical methods for diffusion models on manifolds. In: International Conference on Learning Representations, URL <https://openreview.net/forum?id=PlKWVd2yBkY>
- Lu C, Zhou Y, Bao F, et al (2022) Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. arXiv preprint arXiv:220600927
- Lu C, Zhou Y, Bao F, et al (2023) Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. [2211.01095](#)
- Luo F, Wu X, Guo Y (2024a) And: Adversarial neural degradation for learning blind image super-resolution. Advances in Neural Information Processing Systems 36
- Luo X, Xie Y, Qu Y, et al (2024b) Skipdiff: Adaptive skip diffusion model for high-fidelity perceptual image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 4017–4025
- Mou C, Wu Y, Wang X, et al (2022) Metric learning based interactive modulation for real-world super-resolution. In: European Conference on Computer Vision (ECCV)
- Nan K, Xie R, Zhou P, et al (2024) Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:240702371
- Oquab M, Darabet T, Moutakanni T, et al (2023) Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:230407193
- Park J, Son S, Lee KM (2023) Content-aware local gan for photo-realistic super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10585–10594
- Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763
- Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18, Springer, pp 234–241
- Sahak H, Watson D, Saharia C, et al (2023) Denoising diffusion probabilistic models for robust image super-resolution in the wild. [2302.07864](#)
- Saharia C, Ho J, Chan W, et al (2023) Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(4):4713–4726. <https://doi.org/>

- Sauer A, Karras T, Laine S, et al (2023a) Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In: International conference on machine learning, PMLR, pp 30105–30118
- Sauer A, Lorenz D, Blattmann A, et al (2023b) Adversarial diffusion distillation. [2311.17042](https://arxiv.org/abs/2311.17042)
- Song J, Meng C, Ermon S (2020) Denoising diffusion implicit models. arXiv:201002502 URL <https://arxiv.org/abs/2010.02502>
- Song Y, Sun Z, Yin X (2024) Sdxs: Real-time one-step latent diffusion models with image conditions. arXiv preprint arXiv:240316627
- Sun L, Wu R, Zhang Z, et al (2023) Improving the stability of diffusion models for content consistent super-resolution. arXiv preprint arXiv:240100877
- Tai Y, Yang J, Liu X (2017a) Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3147–3155
- Tai Y, Yang J, Liu X, et al (2017b) Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE international conference on computer vision, pp 4539–4547
- Timofte R, Agustsson E, Van Gool L, et al (2017) Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW
- Wang J, Chan KC, Loy CC (2023a) Exploring clip for assessing the look and feel of images. In: AAAI
- Wang J, Yue Z, Zhou S, et al (2023b) Exploiting diffusion prior for real-world image super-resolution. In: arXiv preprint arXiv:2305.07015
- Wang X, Xie L, Dong C, et al (2021) Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW)
- Wang Y, Yang W, Chen X, et al (2024) Sinsr: diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 25796–25805
- Wang Z, Bovik A, Sheikh H, et al (2004) Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4):600–612. <https://doi.org/10.1109/TIP.2003.819861>
- Wei P, Xie Z, Lu H, et al (2020) Component divide-and-conquer for real-world image super-resolution. In: Proceedings of the European Conference on Computer Vision
- Wu R, Yang T, Sun L, et al (2023) Seesr: Towards semantics-aware real-world image super-resolution. arXiv preprint arXiv:231116518
- Xia B, Zhang Y, Wang S, et al (2023a) Diffir: Efficient diffusion model for image restoration. ICCV
- Xia B, Zhang Y, Wang Y, et al (2023b) Knowledge distillation based degradation estimation for blind super-resolution
- Xie L, Wang X, Chen X, et al (2023) Desra: detect and delete the artifacts of gan-based real-world super-resolution models. arXiv preprint arXiv:230702457
- Yang S, Wu T, Shi S, et al (2022) Maniq: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1191–1200
- Yang T, Ren P, Xie X, et al (2023) Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: arXiv:2308.14469
- Yu F, Gu J, Li Z, et al (2024) Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. arXiv preprint arXiv:240113627

Yue Z, Wang J, Loy CC (2023) Resshift: Efficient diffusion model for image super-resolution by residual shifting. [2307.12348](#)

Zhang A, Yue Z, Pei R, et al (2024a) Degradation-guided one-step image super-resolution with diffusion priors. arXiv preprint arXiv:240917058

Zhang K, Liang J, Van Gool L, et al (2021) Designing a practical degradation model for deep blind image super-resolution. In: IEEE International Conference on Computer Vision, pp 4791–4800

Zhang L, Rao A, Agrawala M (2023a) Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3836–3847

Zhang R, Isola P, Efros AA, et al (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR

Zhang W, Li X, Shi G, et al (2024b) Real-world image super-resolution as multi-task learning. Advances in Neural Information Processing Systems 36

Zhang X, Zeng H, Guo S, et al (2022a) Efficient long-range attention network for image super-resolution. In: European conference on computer vision, Springer, pp 649–667

Zhang Y, Ji B, Hao J, et al (2022b) Perception-distortion balanced admm optimization for single-image super-resolution. In: European Conference on Computer Vision, Springer, pp 108–125

Zhang Y, Huang X, Ma J, et al (2023b) Recognize anything: A strong image tagging model. arXiv preprint arXiv:230603514