

# Unleashing the Power of One-Step Diffusion based Image Super-Resolution via a Large-Scale Diffusion Discriminator

Anonymous Author(s)

Affiliation

Address

email

## 1 Comparison with More Methods

Table 1 presents a quantitative comparison between our  $D^3SR$  and several non-diffusion-based methods, including BSRGAN [1], RealSR-JPEG [2], Real-ESRGAN [3], SwinIR [4], LDL [5], and FeMaSR [6]. Across all evaluated datasets,  $D^3SR$  consistently outperforms these methods on no-reference (NR) IQA metrics. Specifically, while GAN-based methods like BSRGAN and Real-ESRGAN achieve competitive performance on traditional full-reference (FR) metrics such as PSNR and SSIM, they lag behind  $D^3SR$  in NR IQA metrics, which better capture perceptual quality aspects such as image clarity, quality, and detail.

Table 1: Performance comparison of  $D^3SR$  with non-diffusion-model-based NR IQA methods across three datasets. The best results for each metric among the methods are highlighted in red.

Datasets	Methods	PSNR	LPIPS	NIQE	MUSIQ	ManIQA
RealSR	BSRGAN	26.51	0.2685	4.661	63.59	0.5279
	RealSR-JPEG	27.34	0.3962	6.952	36.07	0.3413
	Real-ESRGAN	25.85	0.2729	4.691	59.68	0.5386
	SwinIR	26.43	0.2515	4.686	59.63	0.5111
	LDL	25.53	0.2777	4.691	59.68	0.5386
	FeMaSR	25.43	0.2927	4.686	59.63	0.5111
	$D^3SR$	24.11	0.2961	3.899	68.23	0.6383
RealSet65	BSRGAN	N/A	N/A	4.587	65.58	0.5370
	RealSR-JPEG	N/A	N/A	4.806	50.54	0.4109
	Real-ESRGAN	N/A	N/A	4.408	63.22	0.5497
	SwinIR	N/A	N/A	4.404	63.82	0.5453
	LDL	N/A	N/A	4.676	63.22	0.5395
	FeMaSR	N/A	N/A	4.504	64.88	0.5343
	$D^3SR$	N/A	N/A	3.998	70.25	0.6298

As shown in Table 1,  $D^3SR$  achieves superior performance across all NR IQA metrics compared to non-diffusion-based methods. Specifically,  $D^3SR$  exhibits lower NIQE scores, indicating higher perceptual quality and better image clarity. Additionally,  $D^3SR$  outperforms better in the MUSIQ, ManIQA, and ClipIQA metric, further demonstrating its ability to preserve and enhance image details. These results demonstrate that diffusion-based SR models exhibit stronger generative capability and outperform GAN- or Transformer-based models in the Real-ISR task.

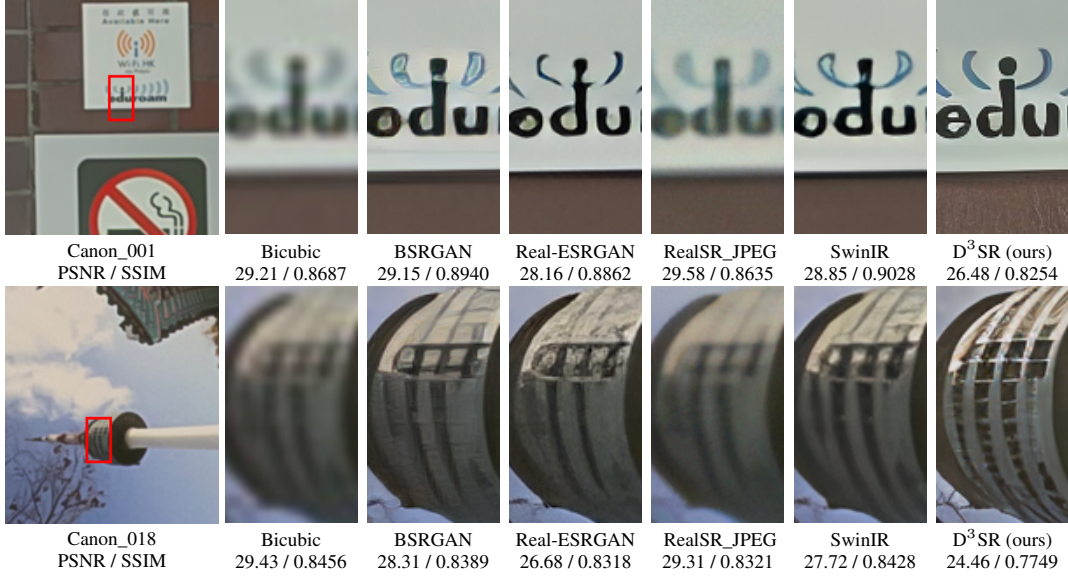


Figure 1: Visual comparison ( $\times 4$ ) of D³SR with GAN-based and Transformer-based methods. Canon\_001 contains the letters ‘edu’. Canon\_018 contains the structures of tower windows. Although GAN-based approaches achieve higher PSNR and SSIM scores, their generated images exhibit less realistic and detailed textures compared to D³SR. Those quantitative and visual comparisons indicate that higher PSNR and SSIM values do not mean better visual quality.

**Visual Comparison.** Figure 1 provides a visual comparison of images generated by D³SR and the non-diffusion-model-based methods mentioned above. While GAN-based methods like BSRGAN and Real-ESRGAN produce visually appealing results, they often introduce artifacts and lack the fine-grained details that D³SR preserves. In contrast, D³SR consistently generates images with sharper edges, more accurate textures, and overall higher visual fidelity, aligning better with human perceptual judgments of image quality.

Despite the competitive performance of GAN-based and transformer-based methods in FR metrics, their NR IQA scores reveal shortcomings in capturing perceptual quality nuances. The superior performance of D³SR in NR IQA metrics indicates its enhanced capability to generate images that are not only quantitatively superior but also qualitatively more pleasing to the human eye. This highlights the importance of incorporating NR IQA evaluations when assessing the true visual effectiveness of super-resolution models.

## 2 Implementation Details

This section provides the implementation details of our D³SR, including model hyperparameters, training procedures, and evaluation settings.

### 2.1 Training Settings

We adopt Stable Diffusion (SD) 2.1-base as the base model of generator for training D³SR, setting both the rank and scaling factor  $\alpha$  of LoRA to 16 in the generator and discriminator. We utilize a learnable embedding as the conditional input for the SD UNet, without any prompts, and remove the text encoder. Training is performed with a batch size of 8 over 100K iterations with 4 NVIDIA A100-40GB GPUs.

During the training process, several key hyperparameters of D³SR are crucial for achieving optimal performance. Table 2 summarizes these important hyperparameters used in our experiments.

Table 2: Key hyperparameters for training D<sup>3</sup>SR.

Hyperparameter	Value
Generator Learning Rate	$5 \times 10^{-5}$
Discriminator Learning Rate	$5 \times 10^{-5}$
Number of Training Iterations	100,000
Batch Size	8
Generator Adversarial Loss Weight ( $\lambda_1$ )	$1 \times 10^{-2}$
EA-DISTS Loss Weight ( $\lambda_2$ )	1

Table 3: Models used for each evaluation metric. All metrics are computed using the pyiqa library. For PSNR and SSIM, evaluations are performed on the Y channel in the YCbCr color space.

Metric	Model File
LPIPS	LPIPS_v0.1_alex-df73285e.pth
DISTS	DISTS_weights-f5e65c96.pth
MUSIQ	musiq_koniq_ckpt-e95806b9.pth
ManIQA	MANIQA_PIPAL-ae6d356b.pth
ClipIQA	RN50.pt (CLIP module)

Table 4: Quantitative comparison ( $\times 4$ ) on valid dataset. Impact of different LoRA rank and  $\alpha$  on D<sup>3</sup>SR performance.

$r$	$\alpha$	NIQE↓	MUSIQ↑	ManIQA↑	ClipIQA↑
4	4	4.0909	68.50	0.6327	0.6521
8	8	3.9706	69.41	0.6365	0.6571
16	16	3.9255	69.21	0.6402	0.6683
8	64	9.4521	29.35	0.3298	0.2808
64	128	5.4717	65.42	0.5853	0.5517

## 2.2 Evaluation Details

We evaluate D<sup>3</sup>SR and other methods on entire images from each test set. Following the implementations of StableSR [7] and OSediff [8], we also apply the Adaptive Instance Normalization (AdaIN) algorithm to post-process generated images, ensuring that the color and style of the generated images closely match those of the input low-resolution (LR) images.

For evaluating large images, we adopt a tiling strategy to address memory limitations. Specifically, each image is divided into overlapping patches of size  $512 \times 512$  pixels, with a 64-pixel overlap between adjacent patches to ensure smooth transitions. We perform inference independently on each image patch and subsequently stitch them together. For the overlapping regions, we average the results to maintain consistency and continuity across the entire image.

The models used for each evaluation metric are listed in Table 3. All metrics are computed using the pyiqa [9] library. For PSNR and SSIM, we evaluate the Y channel in the YCbCr color space of the images to focus on luminance information, which is more indicative of perceived image quality.

## 3 Additional Ablation Studies

In this section, we present further ablation studies that complement those discussed in the main text.

**LoRA Settings.** We primarily investigate the impact of varying the  $\alpha$  and rank settings of LoRA on the performance of D<sup>3</sup>SR. The performance of D<sup>3</sup>SR under different LoRA configurations is presented in Table 4. With lower  $\alpha$  and rank values, the LoRA parameters are insufficient to achieve optimal results. As both  $\alpha$  and rank increase, the fine-tuning capability of LoRA on the model is enhanced, leading to gradual improvements in performance. However, setting either  $\alpha$  or rank too high results in significant overfitting, thereby degrading performance on the test set.

## 59 4 Algorithm of D<sup>3</sup>SR

60 The pseudo-code of our D<sup>3</sup>SR training algorithm is summarized as algorithm 1.

---

### Algorithm 1 Training Algorithm for D<sup>3</sup>SR

---

**Require:**

$\epsilon_\phi$ : Pretrained Stable Diffusion (SD) UNet  
 $E_\phi, D_\phi$ : Pretrained SD VAE Encoder and Decoder  
 $\mathcal{S}$ : Training dataset  
 $N$ : Number of training iterations  
 1: **Initialize** generator  $\mathcal{G}_\theta$  from pretrained SD model:  
      $E_\theta \leftarrow E_\phi$  ▷ Initialize encoder from SD VAE  
      $\epsilon_\theta \leftarrow \epsilon_\phi$  with trainable LoRA ▷ Initialize UNet with LoRA  
      $D_\theta \leftarrow D_\phi$  ▷ Initialize decoder from SD VAE  
 2: **Initialize** guidance module  $\mathcal{D}_\theta$  using downsampling and middle blocks from pretrained SD UNet  
 3: **for**  $i = 1$  **to**  $N$  **do**  
 4:     Sample a batch of  $(x_L, x_H)$  from  $\mathcal{S}$   
    **/\* Generator Step \*/**  
 5:      $z_L = E_\theta(x_L)$  ▷ Encode low-resolution image  
 6:      $\hat{z}_H = \frac{z_L - \sqrt{1 - \bar{\alpha}_{T_L}} \epsilon_\theta(z_L; T_L)}{\sqrt{\bar{\alpha}_{T_L}}}$  ▷ Denoising step  
 7:      $\hat{x}_H = D_\theta(\hat{z}_H)$  ▷ Decode high-resolution image  
 8:      $\mathcal{L}_{\text{spatial}} = L_{\text{MSE}}(x_H, \hat{x}_H) + \lambda_2 L_{\text{EA-DISTS}}(x_H, \hat{x}_H)$   
 9:     Sample  $t \in [0, T]$   
 10:      $\mathcal{L}_\mathcal{G} = -\mathbb{E}_{x_L \sim p_{\text{data}}, t \sim [0, T]} [\log \mathcal{D}_\theta(F(\hat{z}_H, t))]$   
 11:     Update  $\mathcal{G}_\theta$  using  $\mathcal{L}_{\text{spatial}} + \lambda_1 \mathcal{L}_\mathcal{G}$   
    **/\* Discriminator Step \*/**  
 12:      $z_H = E_\theta(x_H)$  ▷ Encode high-resolution image  
 13:     Sample  $t \in [0, T]$   
 14:      $\mathcal{L}_\mathcal{D} = -\mathbb{E}_{x_L \sim p_{\text{data}}, t \sim [0, T]} [\log(1 - \mathcal{D}_\theta(F(\hat{z}_H, t)))]$   
 15:      $\quad -\mathbb{E}_{x_H \sim p_{\text{data}}, t \sim [0, T]} [\log \mathcal{D}_\theta(F(z_H, t))]$   
 16:     Update  $\mathcal{D}_\theta$  using  $\mathcal{L}_\mathcal{D}$   
 17: **return**  $\mathcal{G}_\theta$

---

## 61 5 More Visual Comparisons

62 Figures 2, 3, 4 presents additional visual comparison results with compared methods [7, 10, 11, 12,  
 63 13, 8, 1, 2, 3, 4, 5, 6]. Our D<sup>3</sup>SR demonstrates superior visual quality, detail, and realism in highly  
 64 degraded scenarios, fine hair details, text, and richly textured regions.

## 65 References

- 66 [1] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, “Designing a practical degradation model for  
 67 deep blind image super-resolution,” in *ICCV*, 2021. 1, 4  
 68 [2] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, “Real-world super-resolution via kernel  
 69 estimation and noise injection,” in *CVPRW*, 2020. 1, 4  
 70 [3] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-esrgan: Training real-world blind super-resolution  
 71 with pure synthetic data,” in *ICCV*, 2021. 1, 4  
 72 [4] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration  
 73 using swin transformer,” in *ICCVW*, 2021. 1, 4  
 74 [5] J. Liang, H. Zeng, and L. Zhang, “Details or artifacts: A locally discriminative learning approach  
 75 to realistic image super-resolution,” in *CVPR*, 2022. 1, 4

- 76 [6] C. Chen, X. Shi, Y. Qin, X. Li, X. Han, T. Yang, and S. Guo, “Real-world blind super-resolution  
77 via feature matching with implicit high-resolution priors,” in *ACM MM*, 2022. 1, 4
- 78 [7] J. Wang, Z. Yue, S. Zhou, K. C. K. Chan, and C. C. Loy, “Exploiting diffusion prior for  
79 real-world image super-resolution,” *IJCV*, 2024. 3, 4
- 80 [8] R. Wu, L. Sun, Z. Ma, and L. Zhang, “One-step effective diffusion network for real-world image  
81 super-resolution,” *arXiv preprint arXiv:2406.08177*, 2024. 3, 4
- 82 [9] C. Chen and J. Mo, “IQA-PyTorch: Pytorch toolbox for image quality assessment.” [Online].  
83 Available: <https://github.com/chaofengc/IQA-PyTorch>, 2022. 3
- 84 [10] Z. Yue, J. Wang, and C. C. Loy, “Resshift: Efficient diffusion model for image super-resolution  
85 by residual shifting,” in *NeurIPS*, 2024. 4
- 86 [11] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong, “Diffbir:  
87 Towards blind image restoration with generative diffusion prior,” in *ECCV*, 2024. 4
- 88 [12] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, “Seesr: Towards semantics-aware  
89 real-world image super-resolution,” in *CVPR*, 2024. 4
- 90 [13] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and  
91 B. Wen, “Sinsr: Diffusion-based image super-resolution in a single step,” in *CVPR*, 2024. 4





Figure 2: Visual comparison of super-resolved images generated by D<sup>3</sup>SR and non-diffusion-based methods. D<sup>3</sup>SR produces images with more realistic textures, demonstrating superior perceptual quality.

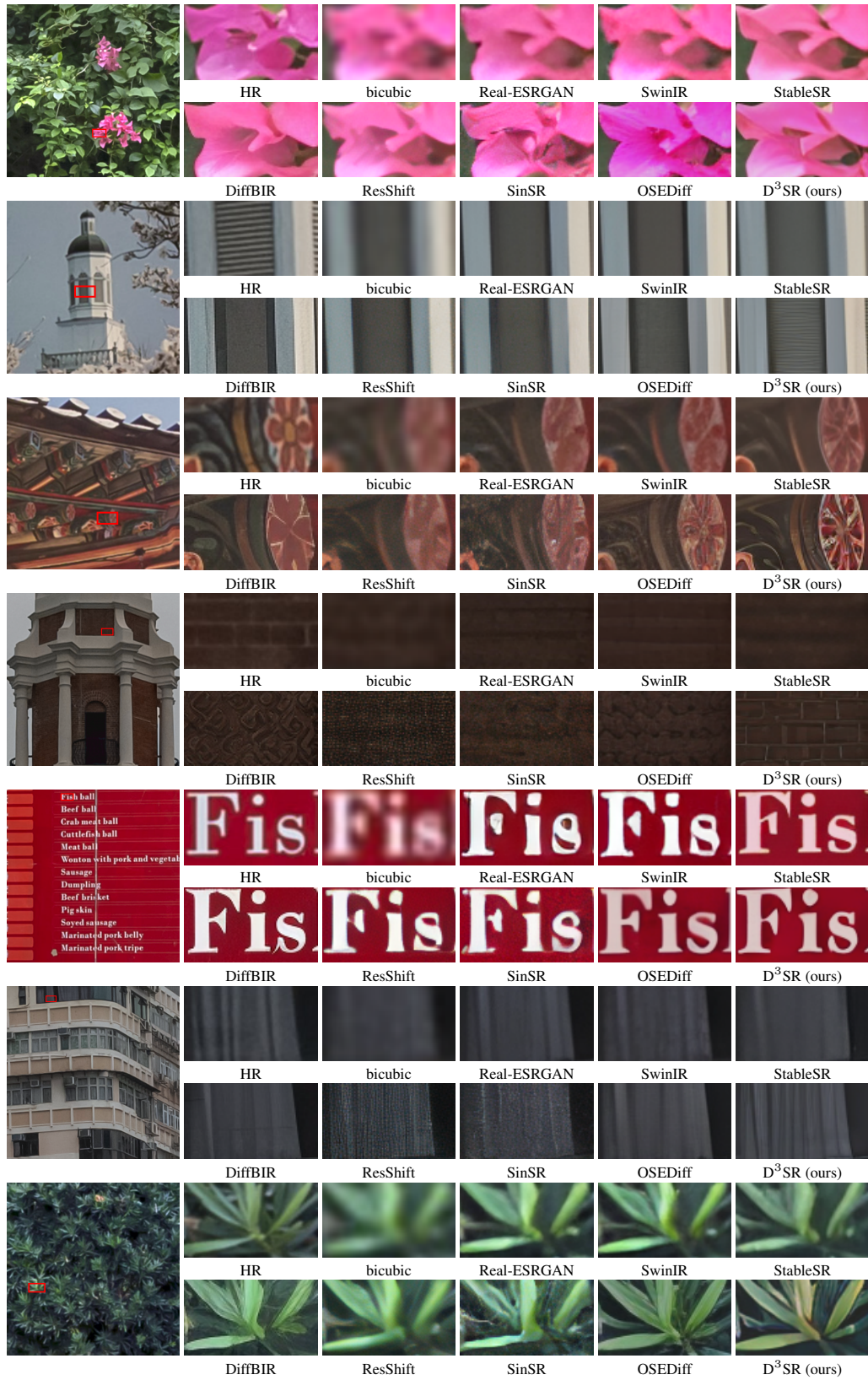


Figure 3: More visualization comparisons of different DM-based Real-ISR methods.





Figure 4: More visualization comparisons of different DM-based Real-ISR methods.