TSD-SR: One-Step Diffusion with Target Score Distillation for Real-World Image Super-Resolution

Supplementary Material

A. Comparison with GAN-based Methods

We compared our method with GAN-based approaches in Tab. 5. While the GAN methods show advantages in full-reference metrics such as PSNR and SSIM, our model outperforms GAN-based methods on all no-reference metrics. Some researchers have found the limitations of PSNR and SSIM in the field of image super-resolution [9, 10]. The effectiveness of PSNR and SSIM in assessing image fidelity in complex degradation scenarios remains debatable, as pixel-level misalignment often arises when restoring severely degraded images. However, no-reference metrics assess image quality based on the individual image, without the need to forcibly align with the ground truth. Therefore, in more complex and realistic degradation scenarios, no-reference metrics may be more suitable for evaluating the results of image super-resolution. In Appendix C, we further discuss the comparison between full-reference metrics and human preferences, and in the Fig. 9, we present a visual comparison with GAN-based methods. From the visualization, it can be observed that our model achieves better results in terms of texture details compared to GAN methods.

B. More Visual Comparisons

In Figs. 10 to 12, we provide more visual caparisons with other diffusion-based methods. Numerous examples demonstrate the robust restoration capabilities of TSD-SR and the high quality of the restored images.

C. Comparisons of Full-reference Metrics and Human Preference

We present additional comparative experiments in Figure 13 to demonstrate that PSNR and SSIM may have limitations in assessing image fidelity under complex degradation scenarios. It can be observed that GAN-based methods with higher PSNR and SSIM produce over-smooth or broken textures, raising concerns about their realism and fidelity. While our approach trades off PSNR and SSIM for natural detail restoration, it achieves enhanced realism and broader perceptual acceptance (Our additional user study reveals that 90.28% of participants prefer ours instead of high PSNR and SSIM methods.). This phenomenon has also been widely discussed in other related research works [1, 3, 7, 9, 10, 12, 13]. LPIPS [12] is proposed to overcome the limitation that PSNR and SSIM fail to align with human judgments in spatial ambiguities situation. Other DMs-based SR researchers [7, 10] argue that DMs introduce superior pre-trained priors, enabling the restoration of information that traditional methods (from scratch) cannot achieve. However, such capability often leads to a decline in pixel-level metrics, as they prioritize distribution modeling and sampling from learned distributions over strict pixel fidelity. We anticipate the development of better full-reference metrics in the future to assess advanced Real-ISR

Table 5. Quantitative comparison with GAN-based methods on both synthetic and real-world benchmarks. The best results of each metric are highlighted in red.

Datasets	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	CLIPIQA ↑
DRealSR	BSRGAN	**28.70**	0.8028	0.2858	0.2143	155.61	6.5408	57.15	0.4847	0.5091
	Real-ESRGAN	28.61	0.8051	0.2818	**0.2088**	147.66	6.7001	54.27	0.4888	0.4521
	LDL	28.20	**0.8124**	**0.2791**	0.2127	155.51	7.1448	53.94	0.4894	0.4476
	FeMASR	26.87	0.7569	0.3156	0.2238	157.72	5.9067	53.70	0.4413	0.5633
	Ours	27.77	0.7559	0.2967	0.2136	**134.98**	**5.9131**	**66.62**	**0.5874**	**0.7344**
RealSR	BSRGAN	26.38	**0.7651**	**0.2656**	0.2123	141.24	5.6431	63.28	0.5425	0.5114
	Real-ESRGAN	**26.65**	0.7603	0.2726	**0.2065**	136.29	5.8471	60.45	0.5507	0.4518
	LDL	25.28	0.7565	0.2750	0.2119	142.74	5.9880	60.92	0.5494	0.4559
	FeMASR	25.06	0.7356	0.2936	0.2285	141.01	5.7696	59.05	0.4872	0.5405
	Ours	24.81	0.7172	0.2743	0.2104	**114.45**	**5.1298**	**71.19**	**0.6347**	**0.7160**
DIV2K-Val	BSRGAN	**24.58**	0.6269	0.3502	0.2280	49.55	4.7501	61.68	0.4979	0.5386
	Real-ESRGAN	24.02	**0.6387**	0.3150	0.2123	38.87	4.8271	60.38	0.5401	0.5251
	LDL	23.83	0.6344	0.3256	0.2227	42.28	4.8555	60.04	0.5328	0.5180
	FeMASR	22.45	0.5858	0.3370	0.2205	41.97	4.8679	57.94	0.4787	0.5769
	Ours	23.02	0.5808	**0.2673**	**0.1821**	**29.16**	**4.3244**	**71.69**	**0.6192**	**0.7416**

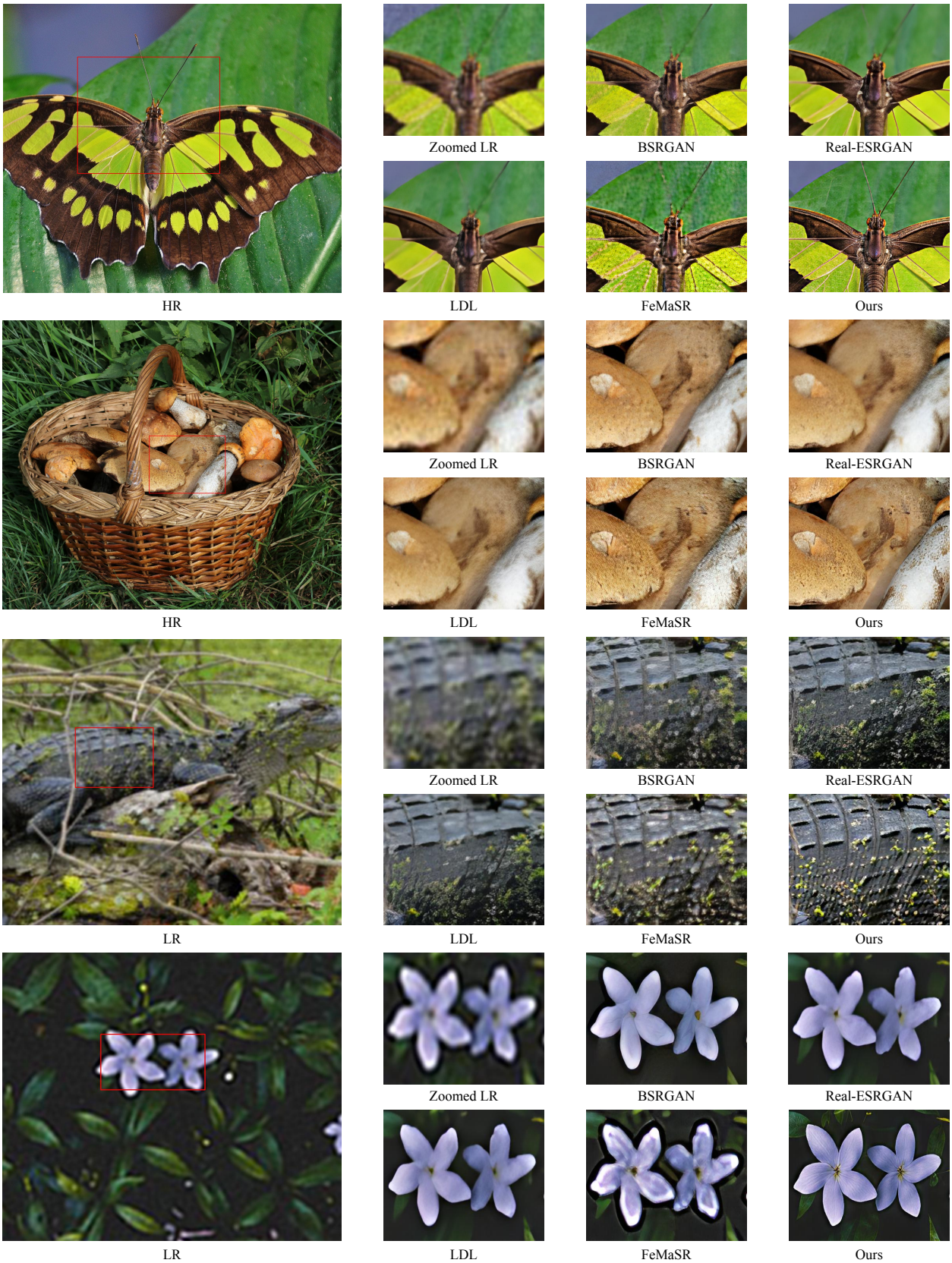

Figure 9. Qualitative comparisons between TSD-SR and GAN-based Real-ISR methods. Please zoom in for a better view.

Table 6. Same base model for fairer comparison.

Model	Base Model	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑	CLIPIQA↑
AddSR	SD2-base	0.3196	0.2242	6.9321	60.85	0.6188
Ours	SD2-base	**0.3040**	**0.2234**	**6.2202**	**65.14**	**0.6935**
OSEDiff	SD2.1-base	0.2968	0.2162	6.4471	64.69	0.6958
Ours	SD2.1-base	**0.2943**	**0.2115**	**5.7934**	**65.41**	**0.7109**

Table 7. Ablation studies for hyperparameter N and s.

N	s	LPIPS↓	DISTS ↓	FID ↓	MUSIQ ↑	MANIQA ↑	CLIPIQA ↑
2	50	0.3104	0.2327	137.64	64.49	0.5717	0.7118
4	**50**	**0.2967**	**0.2136**	**134.98**	**66.62**	**0.5874**	**0.7344**
8	50	0.3421	0.2633	151.64	65.73	**0.5875**	0.7227
4	25	**0.3063**	**0.2201**	**130.74**	**66.19**	0.5828	**0.7269**
4	100	0.3176	0.2230	135.54	65.23	0.5782	0.7024

methods.

D. More Ablation Studies

Base model. To validate the effectiveness of our method across different versions of SD models, we conduct additional experiments, as shown in Tab. 6, including SD2-base and SD2.1-base models. The performance is evaluated on the DRealSR test dataset [8]. It demonstrates superior performance compared to other one-step SR methods, including OSEDiff and AddSR. Specifically, our SD2-base version model outperforms the single-step AddSR across all perceptual reference metrics and no-reference metrics, particularly excelling in NIQE[11], MUSIQ[2], and CLIPIQA[6], significantly surpassing the performance of the AddSR. Our SD2.1-base model demonstrates comparable performance to OSEDiff and surpasses it across various metrics, with notable improvements in NIQE and CLIP-IQA.

Parameters N and s. We provide performance comparisons for different combinations of N and s in Tab. 7. The performance is evaluated on the DRealSR test dataset. N is set to 4, and s is set to 50 in our setting (bold in the table). Larger or smaller N will degrade performance, possibly because it is related to regularization strength. We prefer to use a smaller N because DASM is computationally time-consuming. Therefore, after carefully balancing training duration and model performance, we selected N=4 as the final value. Small s will have similar performance, but the image quality will be compromised when setting large s. Our experimental results indicate that selecting s within the range of 25-75 may yield better performance.

E. Theory of Target Score Matching

The core idea of Target Score Matching (TSM) is that for samples drawn from the same distribution, the real scores predicted by the Teacher Model should be close to each other. Thus we minimize the MSE loss between the Teacher Model's predictions of $\hat{\mathbf{z}}_t$ and $\mathbf{z}_t$ by

$$\begin{aligned}&\mathcal{L}_{\text{MSE}}(\hat{\mathbf{z}}, \mathbf{z}, c_y)\\&= \mathbb{E}_{t,\epsilon}\left[w(t)\|\epsilon_\psi(\hat{\mathbf{z}}_t; t, c_y) - \epsilon_\psi(\mathbf{z}_t; t, c_y)\|_2^2\right]\end{aligned} \tag{10}$$

where the expectation of the gradient is computed across all diffusion timesteps $t \in \{1, \cdots, T\}$ and $\epsilon \sim \mathcal{N}(0, I)$.

To understand the difficulties of this approach, consider the gradient of

$$\nabla_\theta \mathcal{L}_{\text{MSE}}(\hat{\mathbf{z}}, \mathbf{z}, c_y) = \mathbb{E}_{t,\epsilon}\Big[w(t) \cdot \underbrace{\frac{\partial \epsilon_\psi(\hat{\mathbf{z}}_t; t, c_y)}{\partial \hat{\mathbf{z}}_t}}_{\text{Diffusion Jacobian}} \cdot \underbrace{(\epsilon_\psi(\hat{\mathbf{z}}_t; t, c_y) - \epsilon_\psi(\mathbf{z}_t; t, c_y))}_{\text{Prediction Residual}} \cdot \underbrace{\frac{\partial \hat{\mathbf{z}}}{\partial \theta}}_{\text{Generator Jacobian}}\Big] \tag{11}$$

where we absorb $\frac{\partial \hat{\mathbf{z}}_t}{\partial \hat{\mathbf{z}}}$ and the other constant into $w(t)$. The computation of the Diffusion Jacobian term is computationally demanding, as it necessitates backpropagation through the Teacher Model. DreamFusion [4] found that this term struggles with small noise levels due to its training to approximate the scaled Hessian of marginal density. This work also demonstrated that omitting the Diffusion Jaco-

Figure 10. Qualitative comparisons between TSD-SR and different diffusion-based methods. Our method can effectively restore the texture and details of the corresponding object under challenging degradation conditions. Please zoom in for a better view.

Figure 11. Qualitative comparisons between TSD-SR and different diffusion-based methods. Our method can effectively restore the texture and details of the corresponding object under challenging degradation conditions. Please zoom in for a better view.

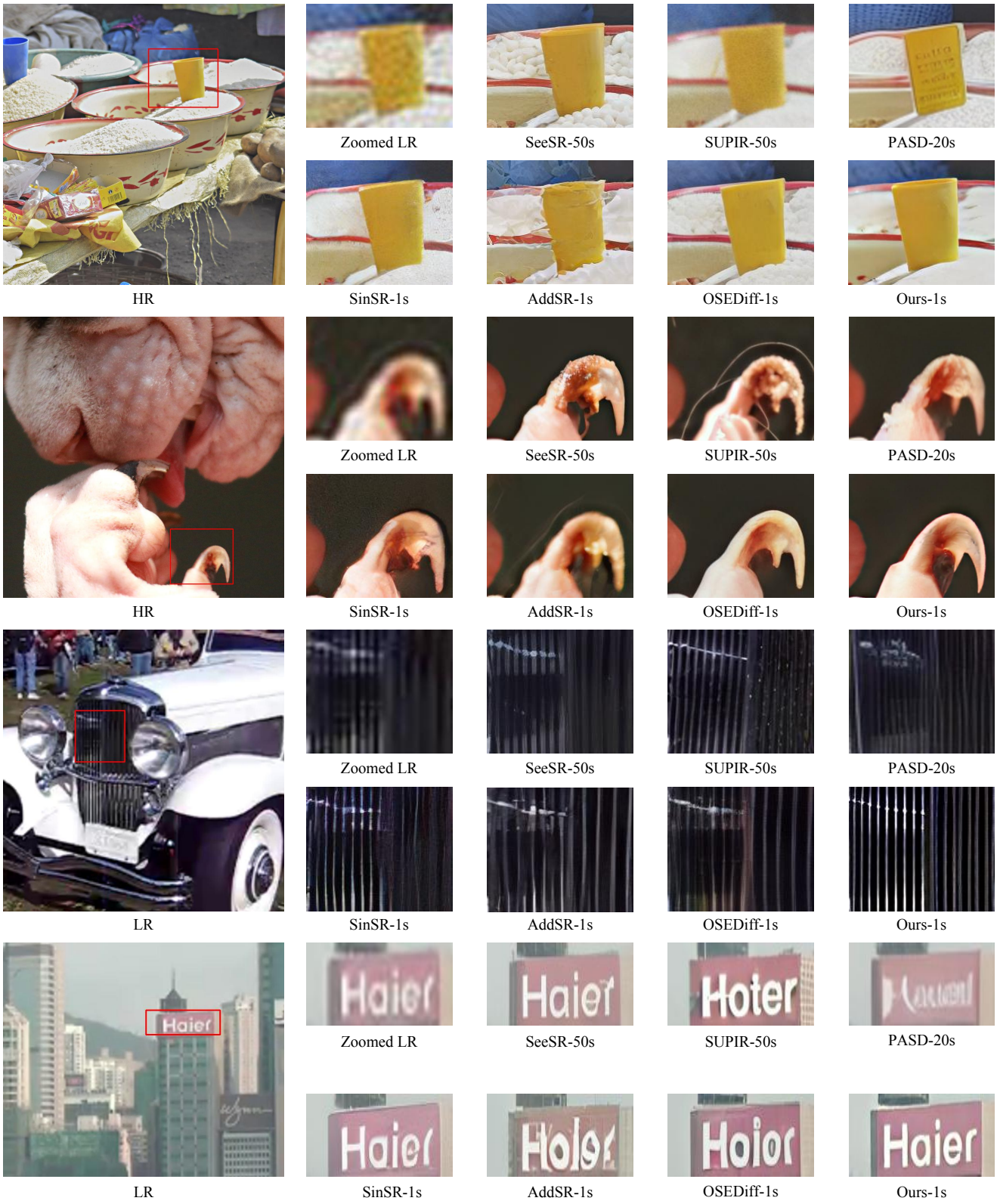

Figure 12. Qualitative comparisons between TSD-SR and different diffusion-based methods. Our method can effectively restore the texture and details of the corresponding object under challenging degradation conditions. Please zoom in for a better view.

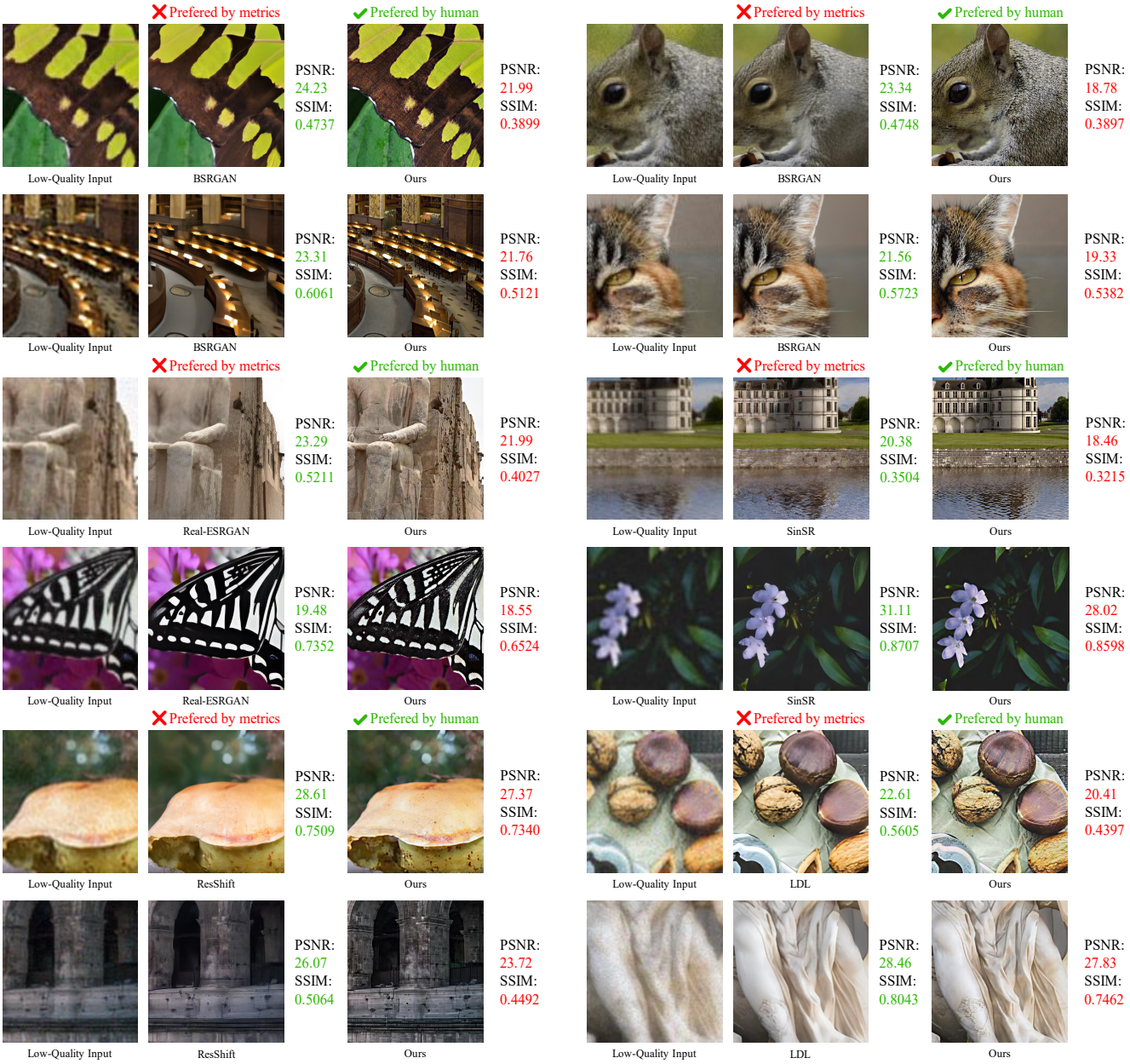

Figure 13. Comparisons between full-reference metric (PSNR/SSIM) assessments and human visual preference. Despite scoring lower on full-reference metrics, TSD-SR generates images that align with human preference.

bian term leads to an effective gradient for optimizing. Similar to their approach, we update Eq. (11) by omitting Diffusion Jacobian:

$$\nabla_\theta \mathcal{L}_{\text{TSM}}(\hat{\mathbf{z}}, \mathbf{z}, c_y) = \mathbb{E}_{t,\epsilon}\Big[w(t) \underbrace{(\epsilon_\psi(\hat{\mathbf{z}}_t; c_y, t) - \epsilon_\psi(\mathbf{z}_t; c_y, t))}_{\text{Prediction Residual}} \underbrace{\frac{\partial \hat{\mathbf{z}}}{\partial \theta}}_{\text{Generator Jacobian}} \Big] \tag{12}$$

The effectiveness of the method can be proven by starting from the KL divergence. We can use a Sticking-the-Landing [5] style gradient by thinking of $\epsilon_\psi(\mathbf{z}_t; c_y, t)$ as a control variate for $\hat{\epsilon}$. For detailed proof, refer to Appendix 4 of DreamFusion [4]. It demonstrates that the gradient of this loss yields the same updates as optimizing the training loss $\mathcal{L}_{\text{MSE}}$ Eq. (10), excluding the Diffusion Jacobian term.

Compared with the VSD loss, we find that the term "Prediction Residual" has changed, and the two losses are similar in the gradient update mode. Specifically, we find that VSD employs identical inputs for both the Teacher and

Algorithm 1: TSD-SR Training Procedure

Input: $\mathcal{D} = \{x_L, x_H, c_y\}$, pre-trained Teacher Diffusion Model including VAE encoder E_ψ, denoising network ϵ_ψ and VAE decoder D_ψ, the number of iterations N and step size s of DASM.

Output: Trained one-step Student Model G_θ.

Initialize Student Model G_θ, including $E_\theta \leftarrow E_\psi$ with trainable LoRA, $\epsilon_\theta \leftarrow \epsilon_\psi$ with trainable LoRA, $D_\theta \leftarrow D_\psi$.
Initialize LoRA diffusion network $\epsilon_\phi \leftarrow \epsilon_\psi$ with trainable LoRA.
while *train* **do**
  Sample $(x_L, x_H, c_y) \sim \mathcal{D}$
  /* Network forward */
  $\hat{\boldsymbol{z}} \leftarrow \epsilon_\theta(E_\theta(x_L))$, $\boldsymbol{z} \leftarrow E_\psi(x_H)$
  $\hat{x}_H \leftarrow D_\psi(\hat{\boldsymbol{z}})$
  /* Compute reconstruction loss */
  $\mathcal{L}_{Rec} \leftarrow LPIPS(\hat{x}_H, x_H)$
  /* Compute regularization loss */
  Sample ϵ from $\mathcal{N}(0, I)$, t from $\{50, \cdots, 950\}$
  $\sigma_t \leftarrow$ FlowMatchingScheduler(t)
  $\hat{\boldsymbol{z}}_t \leftarrow \sigma_t\epsilon + (1-\sigma_t)\hat{\boldsymbol{z}}$, $\boldsymbol{z}_t \leftarrow \sigma_t\epsilon + (1-\sigma_t)\boldsymbol{z}$
  $\mathcal{L}_{Reg} \leftarrow \mathcal{L}_{TSD}(\hat{\boldsymbol{z}}_t, \boldsymbol{z}_t, c_y)$
  for $i \leftarrow 1$ **to** N **do**
    $cur \leftarrow t - i \cdot s$
    $pre \leftarrow t - i \cdot s + s$
    $\sigma_{cur} \leftarrow$ FlowMatchingScheduler(cur)
    $\sigma_{pre} \leftarrow$ FlowMatchingScheduler(pre)
    $\hat{\boldsymbol{z}}_{cur} \leftarrow \hat{\boldsymbol{z}}_{pre} + (\sigma_{cur} - \sigma_{pre}) \cdot \epsilon_\phi(\hat{\boldsymbol{z}}_{pre}; pre, c_y)$
    $\boldsymbol{z}_{cur} \leftarrow \boldsymbol{z}_{pre} + (\sigma_{cur} - \sigma_{pre}) \cdot \epsilon_\psi(\boldsymbol{z}_{pre}; pre, c_y)$
    $\mathcal{L}_{Reg}$ += $weight \cdot \mathcal{L}_{TSD}(\hat{\boldsymbol{z}}_{cur}, \boldsymbol{z}_{cur}, c_y)$
  end
  $\mathcal{L}_G \leftarrow \mathcal{L}_{Rec} + \gamma\mathcal{L}_{Reg}$
  Update θ with $\mathcal{L}_G$
  /* Compute diffusion loss for LoRA Model */
  Sample ϵ from $\mathcal{N}(0, I)$, t from $\{50, \cdots, 950\}$
  $\sigma_t \leftarrow$ FlowMatchingScheduler(t)
  $\hat{\boldsymbol{z}}_t \leftarrow \sigma_t\epsilon + (1-\sigma_t)stopgrad(\hat{\boldsymbol{z}})$
  $\mathcal{L}_{Lora} \leftarrow \mathcal{L}_{Diff}(\hat{\boldsymbol{z}}_t, c_y)$
  Update ϕ with $\mathcal{L}_{Lora}$
end

LoRA models to compute the gradient, while here TSM uses high-quality and suboptimal inputs for the Teacher Model. The losses are related to each other through $\epsilon_\phi(\hat{\mathbf{z}}_t; t, c_y)$.

F. Algorithm

Algorithm 1 details our TSD-SR training procedure. We use classifier-free guidance (cfg) for the Teacher Model and the LoRA Model. The cfg weight is set to 7.5.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 1

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 3

[3] Xiaotong Luo, Yuan Xie, Yanyun Qu, and Yun Fu. Skipdiff: Adaptive skip diffusion model for high-fidelity perceptual image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4017–4025, 2024. 1

[4] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3, 6

[5] Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[6] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 3

[7] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 1

[8] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020. 3

[9] Rui Xie, Ying Tai, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717*, 2024. 1

[10] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. 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*, pages 25669–25680, 2024. 1

[11] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 3

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[13] Yuehan Zhang, Bo Ji, Jia Hao, and Angela Yao. Perception-distortion balanced admm optimization for single-image super-resolution. In *European Conference on Computer Vision*, pages 108–125. Springer, 2022. 1