Supplementary Material to "Pixel-level and Semantic-level Adjustable Super-resolution: A Dual-LoRA Approach"

Lingchen Sun[1,2], Rongyuan Wu[1,2], Zhiyuan Ma[1], Shuaizheng Liu[1,2], Qiaosi Yi[1,2], Lei Zhang[1,2,*]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute
{ling-chen.sun, rong-yuan.wu, zm2354.ma}@connect.polyu.hk
{shuaizhengliu21, qiaosiyijoyies}@gmail.com cslzhang@comp.polyu.edu.hk

The following materials are provided in this supplementary file:

- The detailed derivation from the SD classifier formulation (Eq. (4) in the main paper) to the CSD loss formulation (Eq. (5) in the main paper) (referring to Sec. 3.3 in the main paper);
- Comparisons with GAN-based methods (referring to Sec. 4.1 in the main paper);
- User study (referring to Sec. 4.1 in the main paper.)
- More visual examples of different pixel-semantic scale selections (referring to Sec. 4.2 in the main paper);
- More visual comparisons between our PiSA-SR and DM-based SR methods (referring to Sec. 4.3 in the main paper);
- Ablation studies (referring to Sec. 4.3 in the main paper).

1. The detailed derivation

We rewrite Eq. (4) in the main paper as follows:

$$\nabla \ell_{CSD}^{\lambda_{cfg}} = \mathbb{E}_{t,\epsilon,z_t,c}\left[\frac{w_t}{\sigma_t}(\epsilon_{real}(z_t,t,c) - \epsilon_{real}(z_t,t))\frac{\partial z_H^{sem}}{\partial \theta_{PiSA}}\right], \tag{S1}$$

where the SD model is parameterized by θ, z_H^{sem} is the estimated HQ latent with the PiSA LoRA, c is the text prompt extracted from z_H^{sem}, $\sigma_t = \sqrt{1-\bar{\alpha}_t}$ is the standard deviation, and ϵ_{real} is the output of pre-trained SD model. z_t is obtained by adding noise ϵ to semantic output z_H^{sem} by $z_t = \sqrt{\bar{\alpha}_t}\cdot z_H^{sem} + \sqrt{1-\bar{\alpha}_t}\cdot\epsilon$, where the noise ϵ is sampled from $\epsilon \sim \mathcal{N}(0, I)$. $w_t = \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}} \cdot \frac{CS}{\|f\left(z_t, \epsilon_{real}^{\lambda_{cfg}}\right) - z_H^{sem}\|_1}$ is the timestep-dependent scalar weight introduced in DMD [18] to improve the training dynamics, where S is the number of spatial locations, C is the number of channels, and $f(\cdot)$ is the function defined in Eq. (1) of the main paper, $\epsilon_{real}^{\lambda_{cfg}}$ denotes the pre-trained SD output with the classifier-free guidance (CFG) term ϵ_{real}^{cls} defined in Eq. (6) of the main paper.

Following [14, 18], we calculate the distribution matching gradient within the latent space instead of the noise domain. Therefore, Eq. (S1) can be written as:

$$\nabla \ell_{CSD}^{\lambda_{cfg}} = \mathbb{E}_{t,\epsilon,z_t,c}\left[\frac{w_t\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\left(f\left(z_t, \epsilon_{real}(z_t,t)\right) - f\left(z_t, \epsilon_{real}(z_t,t,c)\right)\right)\frac{\partial z_H^{sem}}{\partial \theta_{PiSA}}\right], \tag{S2}$$

We apply the CFG component to $f\left(z_t, \epsilon_{real}(z_t,t,c)\right)$ in Eq. (S2) and merge the timestep-related weights to obtain Eq. (S4), which is the form of Eq. (5) in the main paper.

$$\nabla \ell_{CSD}^{\lambda_{cfg}} = \mathbb{E}_{t,\epsilon,z_t,c}\left[\frac{CS}{\|f\left(z_t, \epsilon_{real}^{\lambda_{cfg}}\right) - z_H^{sem}\|_1}\left(f\left(z_t, \epsilon_{real}(z_t,t)\right) - f\left(z_t, \epsilon_{real}^{\lambda_{cfg}}\right)\right)\frac{\partial z_H^{sem}}{\partial \theta_{PiSA}}\right]. \tag{S4}$$

*Corresponding author. This work is supported by the PolyU-OPPO Joint Innovative Research Center.

Table S1. Quantitative comparison among the state-of-the-art GAN-based SR methods on synthetic and real-world test datasets. The best results are highlighted in **red**.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
DIV2K	RealESRGAN	24.29	**0.6371**	0.3112	0.2141	37.64	4.68	0.5277	61.06	0.5501
	BSRGAN	**24.58**	0.6269	0.3351	0.2275	44.23	4.75	0.5071	61.20	0.5247
	LDL	23.83	0.6344	0.3256	0.2227	42.29	4.85	0.5180	60.04	0.5350
	PiSA-SR-S1	23.87	0.6058	**0.2823**	**0.1934**	**25.07**	**4.55**	**0.6927**	**69.68**	**0.6400**
RealSR	RealESRGAN	**25.69**	0.7616	0.2727	0.2063	135.18	5.83	0.4449	60.18	0.5487
	BSRGAN	26.39	**0.7654**	**0.2670**	0.2121	141.28	5.66	0.5001	63.21	0.5399
	LDL	25.28	0.7567	0.2766	0.2121	142.71	6.00	0.4477	60.82	0.5485
	PiSA-SR-S1	25.50	**0.7417**	0.2672	**0.2044**	**124.09**	**5.50**	**0.6702**	**70.15**	**0.6560**
DrealSR	RealESRGAN	28.64	0.8053	0.2847	**0.2089**	147.62	6.69	0.4422	54.18	0.4907
	BSRGAN	**28.75**	0.8031	0.2883	0.2142	155.63	6.52	0.4915	57.14	0.4878
	LDL	28.21	**0.8126**	**0.2815**	0.2132	155.53	7.13	0.4310	53.85	0.4914
	PiSA-SR-S1	28.31	0.7804	0.2960	0.2169	**130.61**	**6.20**	**0.6970**	**66.11**	**0.6156**

2. Comparisons with GAN-based methods

We compare PiSA-SR with three representative GAN-based SR methods: RealESRGAN [11], BSRGAN [20], and LDL [7]. The quantitative results are presented in Table S1. PiSA-SR achieves the best performance on no-reference metrics (NIQE [8], CLIPIQA [10], MUSIQ [6], and MANIQA [16]) across the three benchmark datasets [1, 3, 13], due to the enhanced generative capacity of the pre-trained SD model.

In addition, PiSA-SR demonstrates competitive results on reference-based metrics (*e.g.*, LPIPS [21] and DISTS [4]), demonstrating a strong balance between perceptual quality and content fidelity. Fig. S1 provides visual comparisons between PiSA-SR and the GAN-based methods. PiSA-SR generates more realistic details from LQ images, such as the regular textures of rope (see the first group) and the intricate textures of leaves (see the second and third groups).

3. User study

To further validate the effectiveness of the proposed adjustable SR method, we conducted a user study. Each participant was presented with a series of LQ images and their corresponding HQ outputs restored by our method. Specifically, for each LQ input, we applied a fixed pixel-level guidance factor of 1.0, ensuring consistency in pixel-level enhancement. Then, two HQ images were restored using two varying semantic guidance factors of 0.6 and 1.2, respectively. During the study, participants were asked to select one of the two generated HQ images with superior semantic quality. The selection was considered positive if the participant chose the image generated with a semantic guidance factor of 1.2, demonstrating that our method effectively enhances the semantic quality of the output by increasing the semantic guidance factor.

In total, 10 participants provided 500 votes on 50 different LQ images. The results showed a positive selection rate of 98%, which strongly supports the effectiveness of our approach. Therefore, increasing the semantic guidance factor can lead to noticeable improvements in semantic quality, making the generated images more visually appealing and semantically faithful to the input structure.

4. More visual examples on adjustable SR

In Fig. S2, we provide additional visual examples on adjustable SR. The horizontal and vertical axes represent the enhancement scales for semantic and pixel levels, respectively. Increasing pixel-level enhancement gradually reduces degradation and sharpens edges, but excessive enhancement can blur finer details (*e.g.*, the lady's hair accessory in the first group of Fig. S2). On the other hand, raising the semantic-level enhancement improves overall scene details (*e.g.*, the seagull's wings, sea ripples, and island trees in the second group of Fig. S2), but may introduce artifacts when over-enhanced.

5. More visual comparisons with DM-based SR methods

We provide more visual comparisons of DM-based SR methods in Fig. S3. PiSA-SR surpasses other methods by reconstructing more accurate structures (*e.g.*, the red flag's structure in the second group) and producing more realistic details (*e.g.*, the flower's texture in the fourth group and the water ripples in the fifth group of Fig. S3).

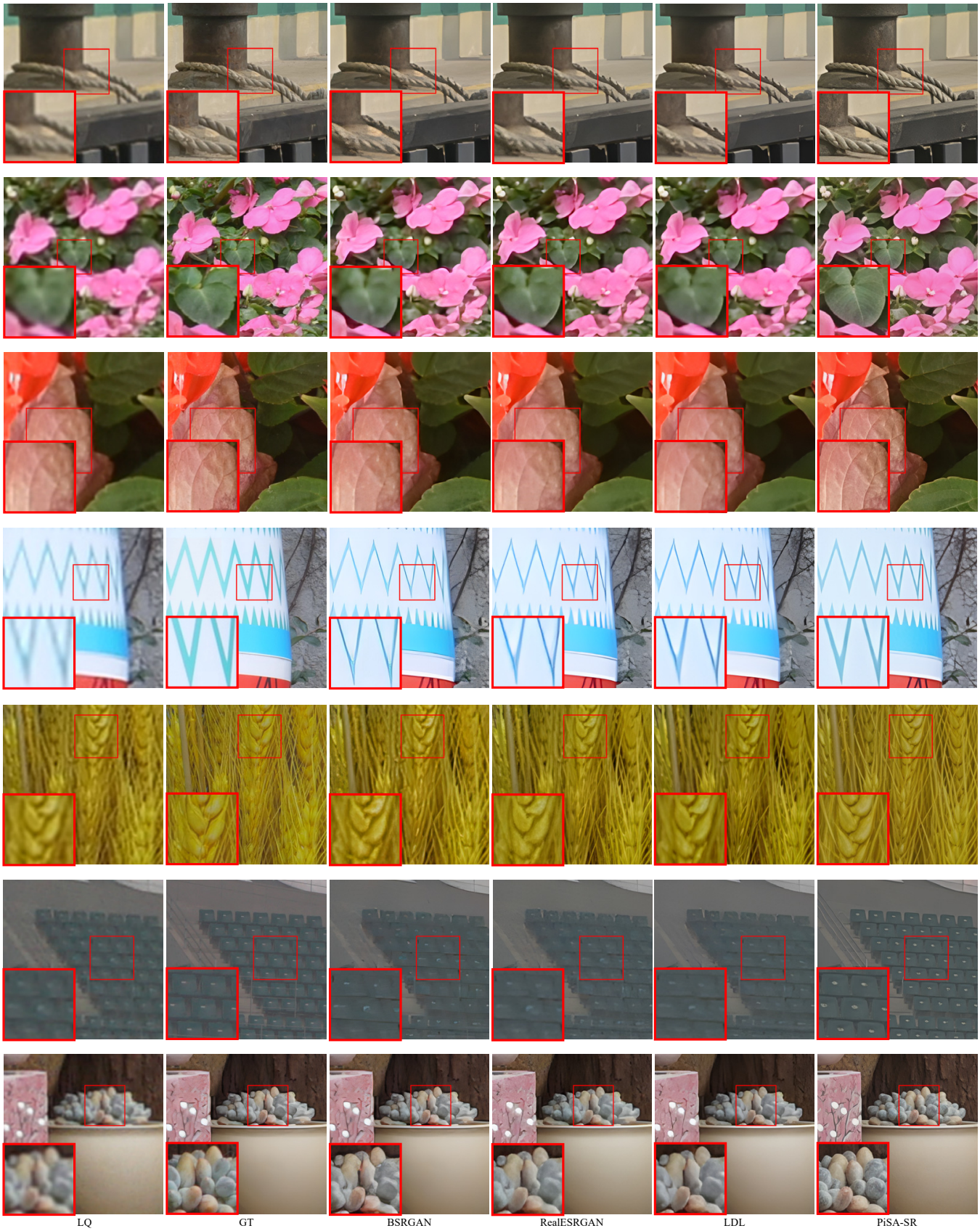


Figure S1. Visual comparisons between PiSA-SR and different GAN-based SR methods. Please zoom in for a better view.

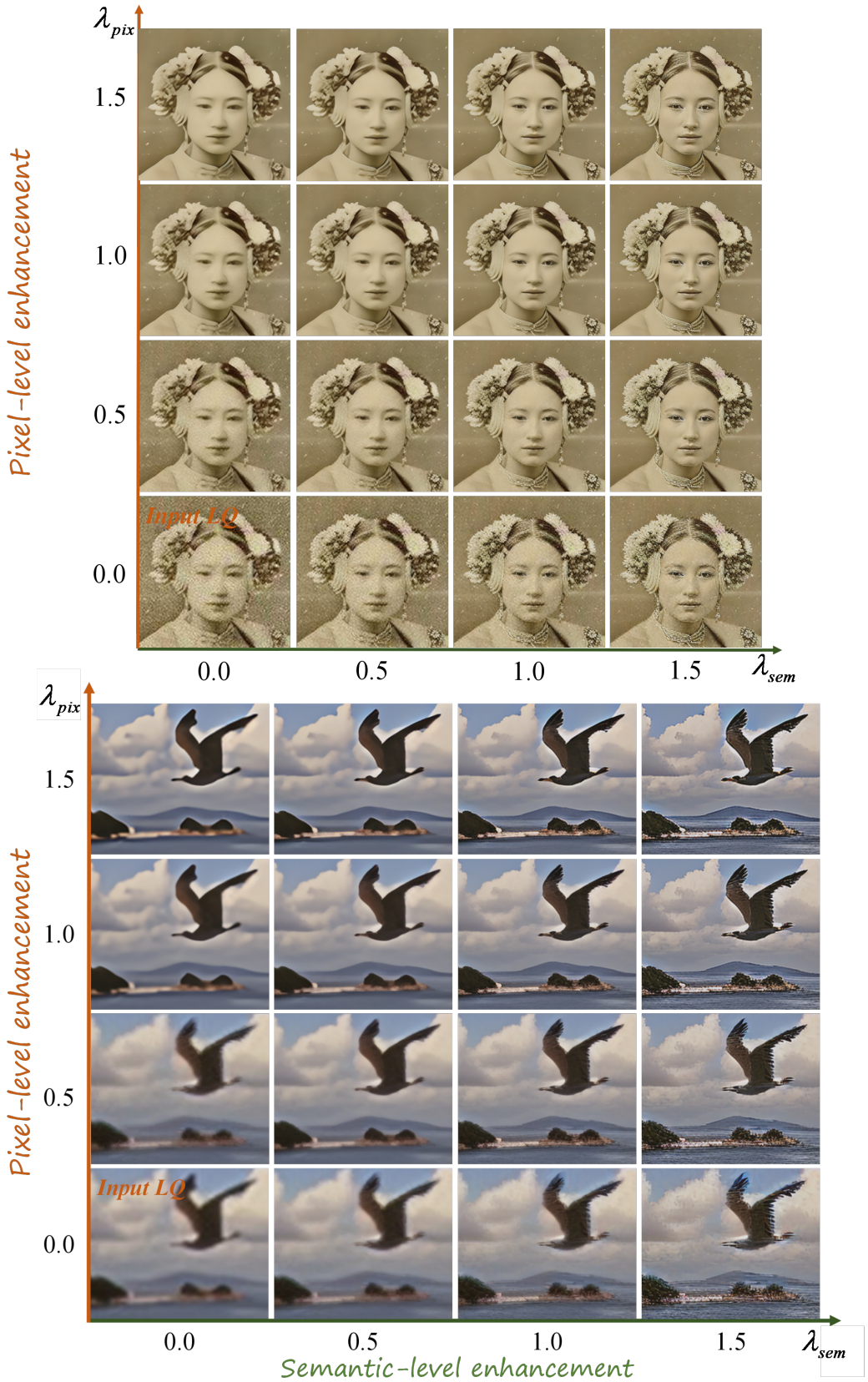


Figure S2. Visual examples of PiSA-SR with different pixel-semantic scales. The horizontal and vertical axes indicate the semantic-level and pixel-level enhancement scales, respectively.

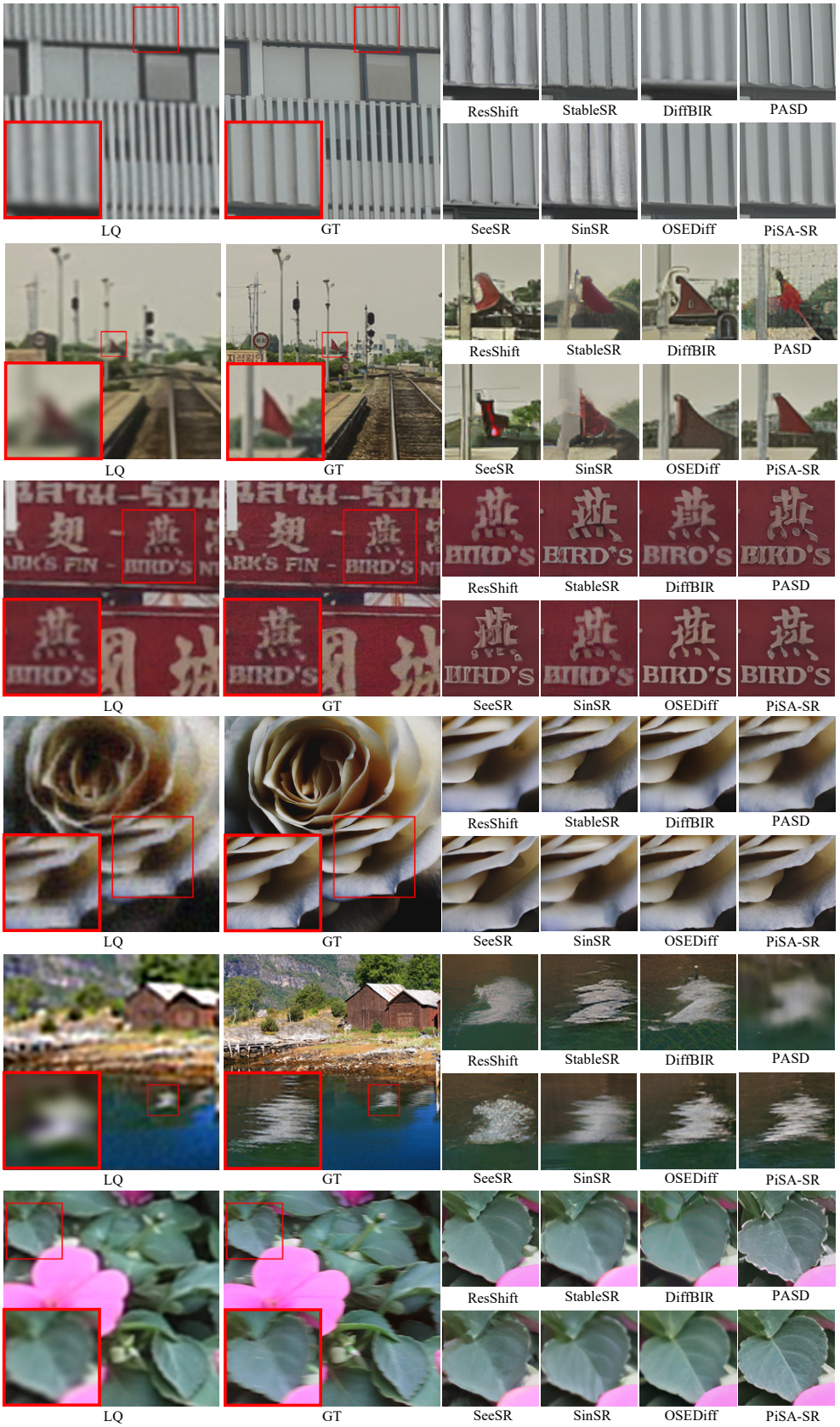


Figure S3. Visual comparisons of different DM-based SR methods. Please zoom in for a better view.

Table S2. Ablation studies on the dual-LoRA training approach on the RealSR dataset.

Methods	Pixel-level LoRA	Semantic-level LoRA	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
V1	✓	×	27.28	0.7975	0.3090	0.3130	0.3995	49.02
V2	×	✓	24.13	0.7290	0.2803	0.6711	0.6614	70.69
PiSA-SR	✓	✓	25.50	0.7417	0.2672	0.6702	0.6560	70.15

Table S3. Comparisons of CFG and the proposed semantic-level guidance on the RealSR dataset.

λ_{cfg}	λ_{sem}	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	Inference time(s)/Image
1.0	1.0	25.50	0.2672	0.6702	70.15	0.09
1.2	×	25.38	0.2684	0.6708	70.23	0.15
1.5	×	25.30	0.2698	0.6708	70.29	0.15
3.0	×	24.61	0.2834	0.6540	70.14	0.15
×	1.2	24.59	0.3000	0.7015	71.60	0.13
×	1.5	23.08	0.3541	0.6835	71.76	0.13

Table S4. Ablation studies on pixel-level and semantic-level LoRA ranks on the RealSR dataset.

Methods	Pixel-level LoRA rank	Semantic-level LoRA rank	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
PiSA-SR	4	4	25.50	0.7417	0.2672	0.6702	0.6560	70.15
R1	4	8	25.40	0.7401	0.2719	0.6734	0.6584	69.93
R2	4	16	25.36	0.7398	0.2726	0.6757	0.6603	70.15
R3	4	32	25.40	0.7394	0.2733	0.6784	0.6634	70.13
R4	8	4	25.39	0.7422	0.2628	0.6663	0.6500	69.86
R5	16	4	25.54	0.7511	0.2624	0.6603	0.6436	69.77
R6	32	4	26.01	0.7565	0.2564	0.6318	0.6409	68.30

6. Ablation studies

6.1. Effectiveness of the dual-LoRA training

To validate the effectiveness of the proposed dual-LoRA training approach, we conduct three experiments, as shown in Table S2: optimizing only the pixel-level LoRA with ℓ_2 loss (V1), optimizing only the semantic-level LoRA with LPIPS and CSD losses (V2), and applying the proposed dual-LoRA training (PiSA-SR). The performance is evaluated on the RealSR test dataset [3]. In V1, the no-reference-based metrics are relatively poor, suggesting that the results lack finer details. In V2, while only optimizing with semantic losses, the reference-based metrics decline, introducing weakened fidelity and unnatural details. In contrast, the proposed dual-LoRA training (PiSA-SR) strikes a better balance between fidelity and perceptual quality. It preserves fine pixel-level details in LQ while enhancing semantic details, resulting in more visually pleasing and natural-looking results.

6.2. Comparison between CFG and the proposed semantic-level guidance

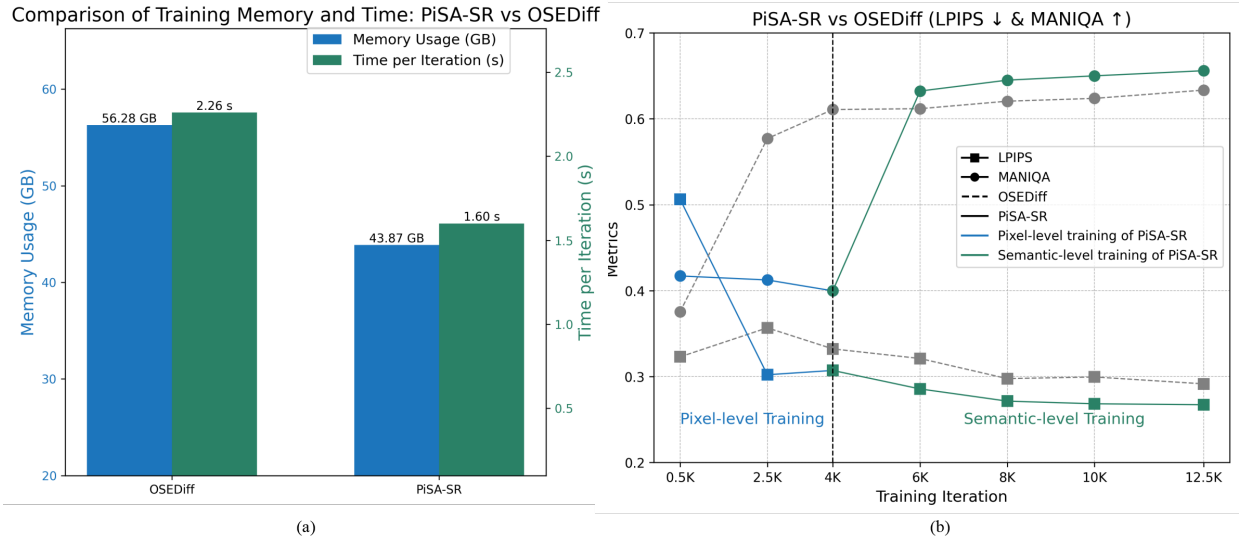
CFG [5] is a commonly used strategy to enhance semantic information in text-to-image (T2I) tasks [9] and multi-step DM-based SR methods [15, 17]. However, it becomes ineffective in one-step DMs distilled from multi-step DMs [14, 18]. As discussed in Sec. 4.2 of the main paper, our proposed semantic-level guidance offers an alternative for one-step DM-based SR methods to semantics enhancement. To further demonstrate its effectiveness, we compare the proposed semantic-level guidance with CFG on the RealSR [3] test dataset in Table S3. Unlike in multi-step DM-based SR methods, increasing the CFG scale in one-step DMs does not improve no-reference metrics (CLIPIQA [10], MUSIQ [6]) and can even degrade both reference- and no-reference-based metrics (*e.g.*, LPIPS [21] and CLIPIQA [10]) at higher scales (*e.g.*, $\lambda_{cfg} = 3.0$). In contrast, adjusting the proposed semantic-level guidance effectively boosts semantic information in the restored image (*e.g.*, changing λ_{sem} from 1.0 to 1.2 can increase the MUSIQ metric from 70.15 to 71.60). Additionally, applying CFG requires more inference time than PiSA-SR, since it needs to extract text prompts from LQ to generate conditional outputs.

6.3. Impact of LoRA rank

In our default setting, the ranks of both pixel-level and semantic-level LoRAs are set to 4. We conduct experiments by fixing one LoRA rank at 4 and varying the other to observe how the performance on the RealSR [3] test dataset changes. The results are shown in Table S4. Increasing the semantic-level LoRA rank enhances semantic details, as reflected in the CLIPIQA and MUSIQ metrics. However, this comes at the cost of image fidelity, resulting in lower reference-based scores (*e.g.*, LPIPS). A

Table S5. Ablation studies on the input timestep on the RealSR dataset.

Methods	Input timestep	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
PiSA-SR	1	25.50	0.7417	0.2672	0.6702	0.6560	70.15
I1	250	25.52	0.7398	0.2685	0.6705	0.6572	70.05
I2	500	25.42	0.7383	0.2699	0.6728	0.6596	70.09
I3	750	25.38	0.7376	0.2701	0.6750	0.6598	70.11
I4	999	25.25	0.7330	0.2707	0.6813	0.6628	70.26



(a) (b)

Figure S4. Comparison between OSEDiff and PiSA-SR: (a) Training memory consumption and time per iteration, and (b) performance (LPIPS [21] and MANIQA [16]) on the RealSR test dataset across training iterations. Experiments are conducted on a single NVIDIA A100 80G GPU with a batch size of 4.

similar pattern can be observed for pixel-level LoRA. Raising its rank improves fidelity, as shown by the SSIM and LPIPS, but reduces some image details, leading to a drop in no-reference metrics (*e.g.*, MUSIQ). Overall, increasing the LoRA rank for one task (pixel-level or semantic-level enhancement) improves its performance but deteriorates the other. This is due to the inherent conflict between these two tasks [2]. We choose a rank of 4 for both LoRA modules as it offers a good balance.

6.4. Impact of the input timestep

In one-step DM-based SR methods, the input timestep must be set in advance. By default, we use a value of 1. Table S5 presents the performance of different timestep configurations on the RealSR [3] test dataset. Generally speaking, increasing the input timestep enhances the semantic details in the restored images but worsens the fidelity performance. This is because the pre-trained models with larger input timestep have stronger denoising capabilities, which improve the generation capability but reduce the ability to preserve the input structures. In PiSA-SR, we set the input timestep to 1 to retain the LQ image information as much as possible.

6.5. Comparison with OSEDiff

In Sec. 4.3 of the main paper, we present both quantitative and qualitative comparisons between PiSA-SR and OSEDiff [14]. Here, we compare them in three additional aspects: training memory usage, training time per iteration, and convergence speed. PiSA-SR uses the CSD loss [19] for semantic enhancement, while OSEDiff employs the VSD loss [12] for distillation. As discussed in Sec. 3.3 of the main paper, optimizing with the CSD loss provides richer semantic information for the SR model than VSD. Furthermore, CSD does not require bi-level optimization, significantly reducing training memory usage and training time per iteration, as shown in Fig. S4 (a). PiSA-SR formulates the SR task by learning the residual between LQ and HQ latent features (see Sec. 3.1 of the main paper), accelerating training convergence as evidenced in Fig. S5, where PiSA-SR restores clearer text faster than OSEDiff during the first 2000 training iterations.

We also compare the behaviors of PiSA-SR and OSEDiff over the training process, measured by the reference-based

Figure S5. Visual comparisons of SR results from OSEDiff and PiSA-SR across 1 to 2000 training iterations.

LPIPS and no-reference-based MANIQA metrics, as shown in Fig. S4 (b). In the first 4K iterations, PiSA-SR focuses on pixel-level LoRA optimization, improving LPIPS but slightly worsening MANIQA. Because OSEDiff applies LPIPS and VSD losses throughout the training process, it initially outperforms PiSA-SR in MANIQA. After 4K iterations, PiSA-SR begins optimizing the semantic-level LoRA. Thanks to its faster convergence, PiSA-SR outperforms OSEDiff in LPIPS and MANIQA between 4K and 6K iterations. Beyond 6K iterations, both methods continue to improve, but PiSA-SR consistently demonstrates superior performance throughout the training process. Overall, PiSA-SR achieves faster training speeds and lower memory consumption than OSEDiff, despite utilizing pixel-level and semantic-level optimizations.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 2

[2] 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. 7

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 6, 7

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 2

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[6] 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. 2, 6

[7] Jie Liang, Hui Zeng, and Lei Zhang. Details or Artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 2

[8] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[10] 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. 2, 6

[11] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 2

[12] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 7

[13] 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. 2

[14] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 1, 6, 7

[15] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 6

[16] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 2, 7

[17] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 6

[18] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024. 1, 6

[19] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and XIAOJUAN QI. Text-to-3d with classifier score distillation. In *The Twelfth International Conference on Learning Representations*, 2024. 7

[20] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[21] 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. 2, 6, 7