Bridging Fidelity-Reality with Controllable One-Step Diffusion for Image Super-Resolution - Supplemental Material -

Hao Chen Junyang Chen Jinshan Pan Jiangxin Dong*
School of Computer Science and Engineering, Nanjing University of Science and Technology
https://github.com/Chanson94/CODSR

In this supplemental material, we first provide a more comprehensive comparison with GAN-based methods in Section A. We then discuss about the complexity of the proposed CODSR in Section B. Additionally, we provide more analyses of TMG to demonstrate its effectiveness in Section C. Section D shows more diffusion attentive attribution maps (DAAMs) [8]. We also present human evaluation and downstream task validation in section E, followed by more visual comparisons against other methods in Section F.

A. Comparison with GAN-based Methods

We compare the proposed approach with two representative GAN-based methods, including Real-ESRGAN [10] and BSRGAN [17]. Table 6 shows the quantitative results. Although GAN-based methods perform best in terms of PSNR and SSIM, they exhibit poor performance in perceptual-oriented metrics. Our method outperforms all competing approaches across all no-reference metrics. Figure 8 shows visual comparisons on RealSR [2] and DRealSR [12] benchmarks. The results generated by GAN-based methods exhibit perceptually unpleasant artifacts as shown in Figure 8(a1) and Figure 8(b1). In contrast, our method shows realistic structures and details.

Table 6. Quantitative comparison with GAN-based methods on three real-world test datasets. ↑ indicates higher is better, ↓ indicates lower is better. The best and the second-best results are highlighted in **red** and blue, respectively.

Datasets	Metrics	PSNR (dB)↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA+↑
RealSR	Real-ESRGAN [10]	25.69	0.7614	0.2709	0.2061	5.80	60.37	0.5492	0.4323
	BSRGAN [17]	**26.38**	**0.7651**	**0.2656**	0.2124	5.64	63.28	0.5416	0.4576
	Ours	25.37	0.7284	0.2741	**0.2002**	**5.31**	**70.54**	**0.6727**	**0.5768**
DRealSR	Real-ESRGAN [10]	28.61	**0.8051**	**0.2819**	**0.2089**	6.70	54.28	0.4900	0.4088
	BSRGAN [17]	**28.70**	0.8028	0.2858	0.2144	6.54	57.16	0.4855	0.4174
	Ours	28.19	0.7761	0.2919	0.2108	**5.97**	**67.05**	**0.6278**	**0.5589**
RealPhoto60	Real-ESRGAN [10]	-	-	-	-	3.93	59.28	0.5079	0.4395
	BSRGAN [17]	-	-	-	-	5.38	45.46	0.3719	0.3397
	Ours	-	-	-	-	**3.42**	**72.72**	**0.6255**	**0.6005**

Table 7. Run-time performance and model complexity of one-step DM-based SR methods for the diffusion process. All methods are tested on a machine with a NVIDIA 4090 GPU. The best and the second-best results are highlighted in red and blue, respectively.

Methods	OSEDiff [13]	PiSA-SR [7]	TVT [15]	HYPIR [5]	Ours
Run-time (s)	0.39	0.36	2.09	1.50	0.49
#GPU Mem (M)	5915	6614	10394	12835	5942
#Params (M)	1765	1290	1726	1549	1768

*Corresponding author

Figure 8. Qualitative comparisons between CODSR and GAN-based methods on the RealSR [2] and DrealSR [12] datasets.

B. Complexity Analysis

Benefiting from the lightweight design of LQFM, which modulates only the features of the first convolutional layer in the U-Net, our method introduces merely the runtime of two additional MLP layers compared with SD 2.1-base [6]. We further conduct a run-time benchmark on the diffusion process of existing one-step DM-based methods on 100 images, where the results are averaged over 100 images with the resolution of 1024×1024 pixels. As shown in Table 7, our approach exhibits competitive inference efficiency. In addition, the Table 8 below details the inference latency and computational costs of the proposed submodules, showing that our method incurs manageable training costs.

Table 8. Effectiveness and cost of each module in our proposed network.

	RGPA	LQFM	TMG	DrealSR					Training		Inference	
				PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA+↑	Time(h)	Mem(G)	Latency(s)	#FLOPs(T)
Base	✗	✗	✗	28.31	0.7712	0.2906	65.89	0.5038	7.5	22.07	0.39	10.19
$\text{Base}_{\text{w/ RGPA}}$	✓	✗	✗	28.24	0.7699	0.2914	66.26	0.5109	7.6	22.07	0.39	10.19
$\text{Base}_{\text{w/ LQFM}}$	✗	✓	✗	**28.54**	0.7754	**0.2876**	65.90	0.5106	7.6	22.15	0.46	10.24
$\text{Base}_{\text{w/ RGPA\&LQFM}}$	✓	✓	✗	28.43	0.7724	0.2902	66.27	0.5213	7.7	22.15	0.46	10.24
Ours	✓	✓	✓	28.19	**0.7761**	0.2919	**67.05**	**0.5589**	10.4	23.89	0.49	10.29

C. More Detailed Analyses of TMG

To further investigate the influence of semantic knowledge distillation and interactive alignment on generative capability, we compare our CODSR with three baselines that respectively *(i)* removes the text-matching guidance and replaces the VSD loss with the GAN loss, *(ii)* removes the text-matching guidance, and *(iii)* replaces the VSD loss with the GAN loss in the second stage. As shown in Table 9, using the GAN loss with the text-matching guidance fails to fully unleash the semantic generation capability due to the lack of effective semantic guidance in the second-stage training. Merely employing the VSD loss without the text-matching guidance enables semantic knowledge distillation from the pretrained model, to some extent. Our proposed TMG further improves the generation quality by achieving precise text–image interaction through explicit spatial guidance.

Table 9. Comparison of different semantic enhancement strategies on the DrealSR [12] benchmark.

	NIQE ↓	MUSIQ ↑	MANIQA ↑	CLIPIQA+ ↑
$\text{Base}_{\text{w/ RGPA\&LQFM}}$	6.00	66.27	**0.6317**	0.5213
(i)	6.07	65.56	0.6257	0.5009
(ii)	6.02	66.70	0.6253	0.5487
(iii)	6.08	65.21	0.6259	0.5015
Ours	**5.97**	**67.05**	0.6278	**0.5589**

D. More Visualization of DAAMs

We present more diffusion attentive attribution maps (DAAMs) [8] for different scenes. Figure 9 shows that when existing diffusion-based methods struggle to extract accurate structural information from LQ images, the text embeddings may generate low responses in the LQ features, hindering the ability of diffusion priors to restore faithful details. In contrast, our proposed TMG helps activate concepts that exhibit weak or negligible responses in standard cross-attention, thereby enabling the recovery of richer and more realistic details and producing a more plausible DAAM [8].

E. Human Evaluation and Downstream Task Validation

As suggested, we further evaluate our method by a user study and the OCR recognition recall on the Occluded RoadText 2024 dataset [9]. As shown in Table 10, our method demonstrates a clear and substantial dominance in user preference, achieving 75.6% of user votes. In addition, the OCR recognition recall of our restored images is better.

Table 10. User study and OCR recognition results of one-step DM-based SR methods. The best and the second-best results are highlighted in red and blue, respectively.

Methods	OSEDiff	PiSA-SR	TVT	HYPIR	Ours
User Preference (%)	6.7	1.1	2.8	13.9	75.6
OCR Recall (%)	40.3	39.3	36.1	38.9	42.7

F. More Qualitative Comparisions

In this section, we present additional visual comparisons with state-of-the-art methods [4, 5, 7, 10, 11, 13–16] on real-world benchmarks. As shown in Figures 10-12, our proposed method can recover more faithful structural details.

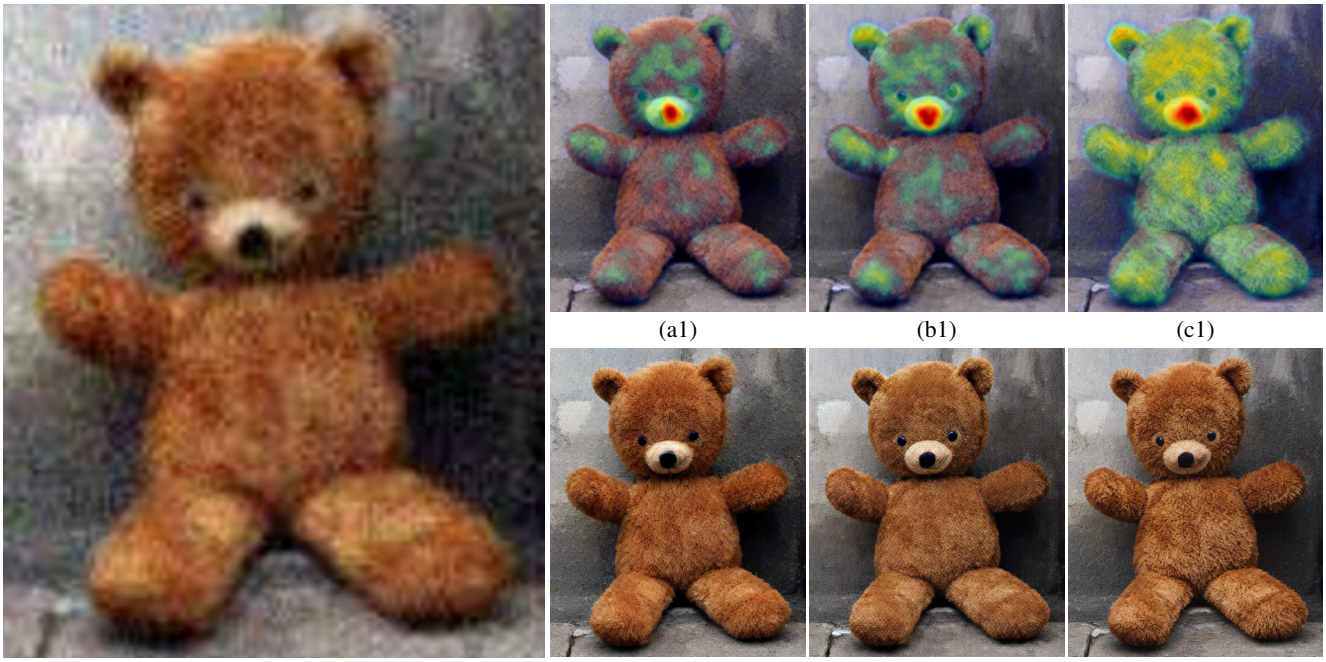

Figure 9. Visualization comparison of DAAMs [8] for the query word *"teddy"*. (a1)-(c1) are DAAMs for OSEDiff [13], TVT [15], and our method. Compared to OSEDiff [13] and TVT [15], our method achieves more precise text–image interaction in the semantic region corresponding to the teddy, resulting in more realistic results. Please zoom in for a better view.

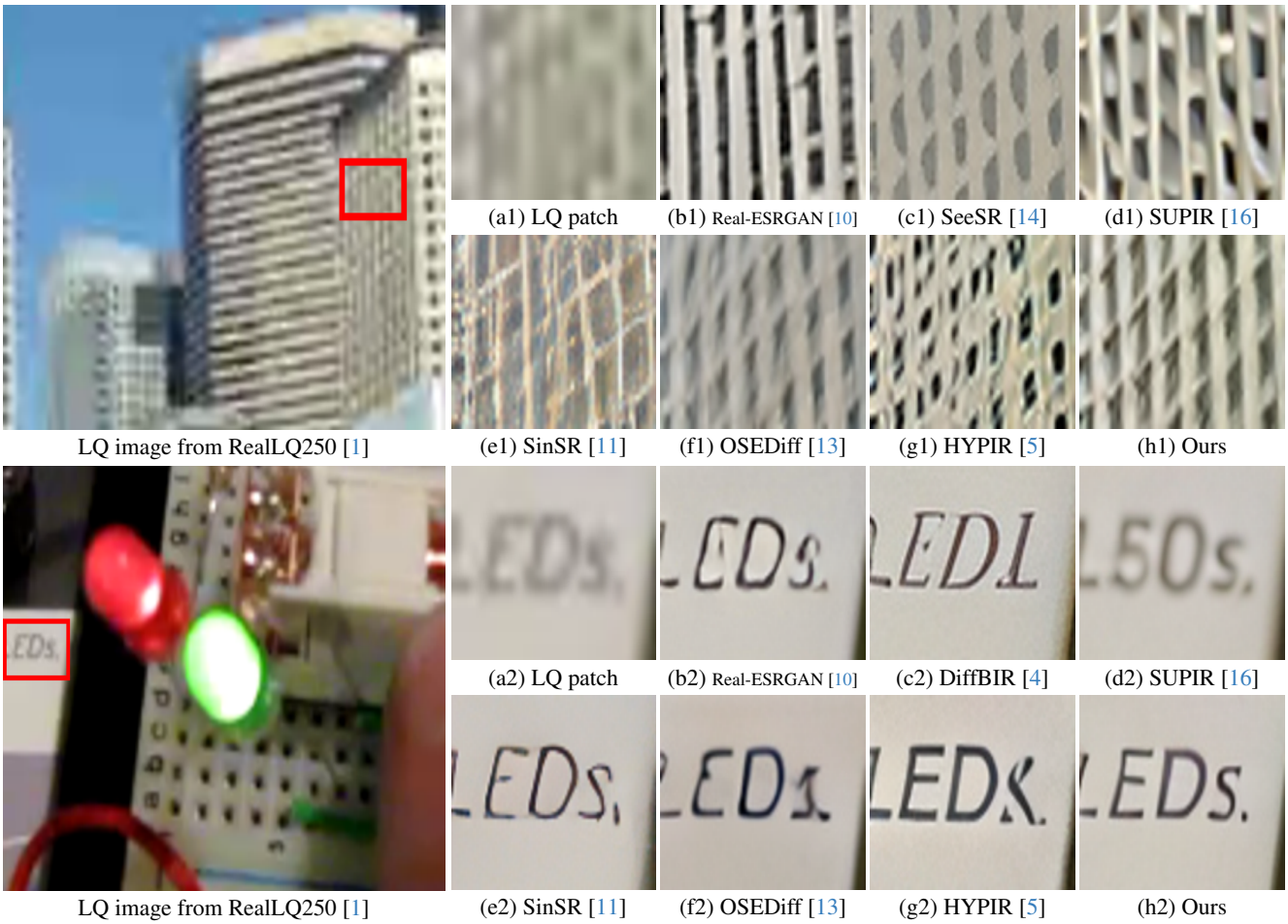

Figure 10. Qualitative comparisons of different methods on the RealLQ250 [1] dataset. Please zoom in for a better view.

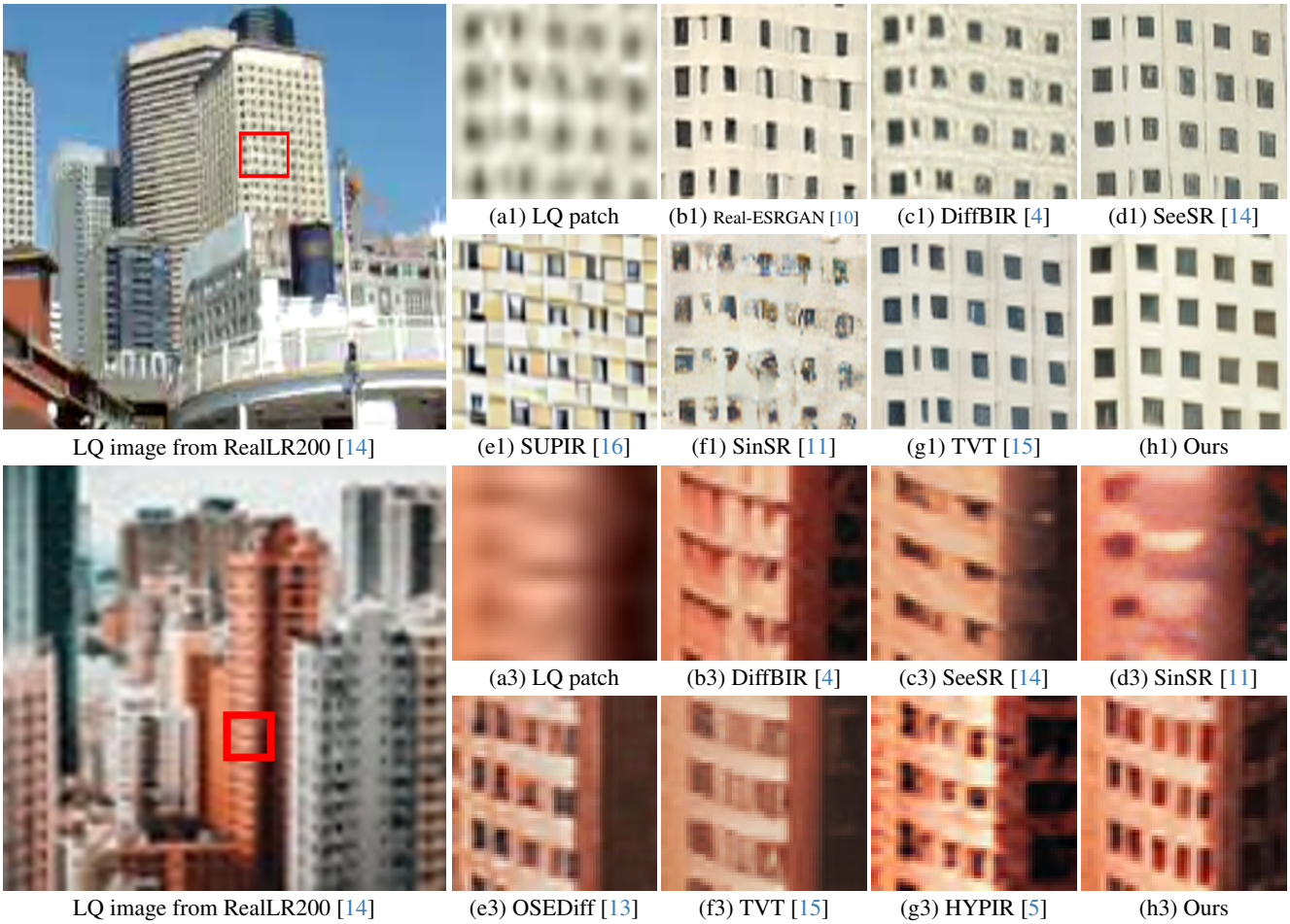

Figure 11. Qualitative comparisons of different methods on the RealLR200 [14] dataset. Please zoom in for a better view.

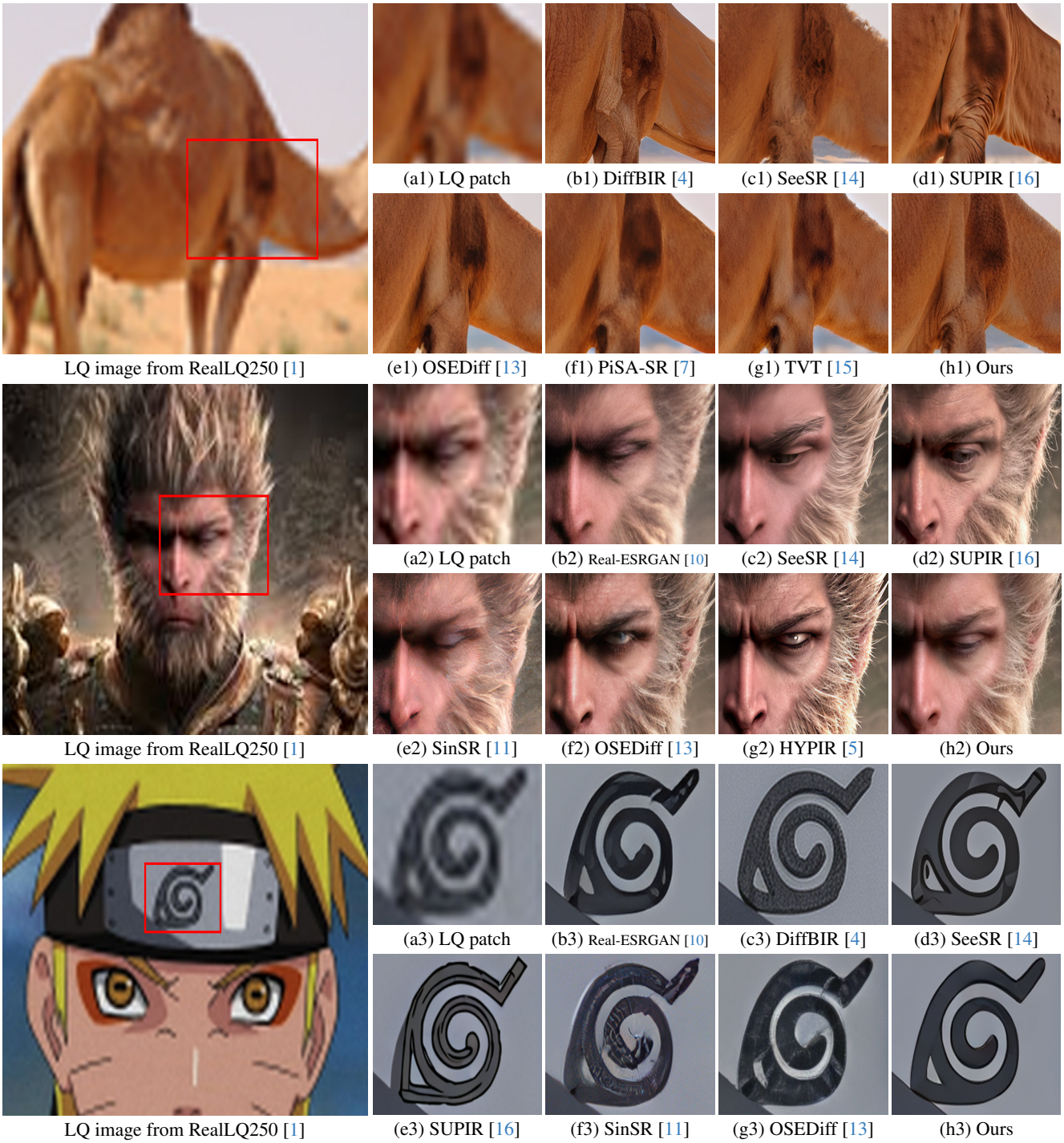

Figure 12. Qualitative comparisons of different methods on the RealLQ250 [1] dataset. Please zoom in for a better view.

References

[1] Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *NeurIPS*, 37:55443–55469, 2024. 5, 7

[2] 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 *ICCV*, 2019. 1, 2

[3] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPR*, 2023. 4

[4] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, 2024. 3, 5, 6, 7

[5] Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *arXiv preprint arXiv:2507.20590*, 2025. 1, 3, 5, 6, 7

[6] Stability.ai. Sd, 2021. https://stability.ai/stablediffusion. 2

[7] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *CVPR*, 2025. 1, 3, 7

[8] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022. 1, 3, 4

[9] George Tom, Minesh Mathew, Ajoy Mondal, Dimosthenis Karatzas, C. V. Jawahar, and Jerod Weinman. Icdar2024 challenge on occluded roadtext. In *ICDAR2024 Workshops*, 2024. 3

[10] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 1, 2, 3, 5, 6, 7

[11] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024. 3, 5, 6, 7

[12] 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 *ECCV*, 2020. 1, 2, 3

[13] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024. 1, 3, 4, 5, 6, 7

[14] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 5, 6, 7

[15] Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer vae training. *arXiv preprint arXiv:2507.20291*, 2025. 1, 4, 6, 7

[16] 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 *CVPR*, 2024. 3, 5, 6, 7

[17] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 1, 2