CASR: A Robust Cyclic Framework for Arbitrary Large-Scale Super-Resolution with Distribution Alignment and Self-Similarity Awareness

Wenhao Guo[1], Zhaoran Zhao[1], Peng Lu[1,*], Sheng Li[2], Qian Qiao[1], RuiDe Li[1]

[1]Beijing University of Posts and Telecommunications [2]Peking University

{whguo, zhaozhaoran, lupeng, qqiao, deruili}@bupt.edu.cn, lisheng@pku.edu.cn

1. Training Details

The GAN loss L_{GAN} employs a vision-aided discriminator [9] with a DINO backbone. The loss weights are set as: $\lambda_1 = 2.0$ (L2), $\lambda_2 = 5.0$ (LPIPS), $\lambda_3 = 0.5$ (GAN), $\lambda_4 = 1.0$ (L_{depth}), and $\lambda_5 = 1.0$ (L_{corr}). Training images are cropped to 512×512 for the first stage and 1024×1024 for the second stage. LR inputs are generated by bicubic downsampling with random scale factors between $\times 1$ and $\times 4$. Data augmentation includes random horizontal and vertical flips.

2. Inference Time Evaluation

We evaluate the inference time and model parameters of various ASISR methods using input images of resolution 128×128 with a $\times 4$ upsampling factor. All experiments are conducted on an NVIDIA A6000 GPU. The reported time for CASR (0.59s) reflects the total three-step inference ($\times 4 \times 3 \times 1.5$). The results are summarized in Table 1. CASR achieves superior inference speed compared to other diffusion-based methods, attributed to the lightweight preprocessor and the one-step SD-Turbo architecture.

Table 1. Comparison of inference time across different models.

Models	LINF	BFSR	IDM	Kim	LIIF+Diff	CiaoSR+Diff	Ours
Time (s)	0.49	0.11	21.3	7.09	3.59	4.47	0.59
Param (M)	17.5	22.1	116.6	157.9	383.62	384.7	529.04

3. Additional Comparisons on Real-World Super-Resolution Datasets

We provide additional comparisons on real-world super-resolution datasets. Three types of datasets are included: (1) Benchmark, containing high-quality images from Set5 [1], Set14 [14], BSD100 [10], and Urban100 [7]. These images remain clean and free from artificial degradations. (2) RealSRSet [15], a collection of web-crawled images exhibiting diverse and complex degradations such as blur, compression artifacts, and noise. (3) COZ (Continuous Optical Zooming) [5], which captures real-world images at varying zoom levels and serves as a challenging benchmark for generalization. Quantitative results are reported in Tables 2 and 3, showing that our method consistently achieves superior perceptual performance across all datasets under various large-scale settings.

4. Additional Visual Results

We provide additional qualitative comparisons to further demonstrate the effectiveness of the proposed method in large-scale super-resolution, as shown in Fig. 1.

5. Distortion Metrics on DIV8K

We supplement the PSNR and SSIM fidelity metrics on DIV8K across all five scales in Table 4. Among generative ASISR methods (IDM, Kim), CASR achieves the highest PSNR/SSIM at $\times 12$ and above, demonstrating that the cyclic framework preserves structural fidelity even at extreme magnifications. Note that INR-based methods (LINF, BFSR) and hybrid pipelines (LIIF+Diff, CiaoSR+Diff) attain higher PSNR/SSIM by design, as they favor pixel-wise regression over perceptual realism. As shown in Table 1 of the main paper, CASR substantially outperforms all methods on perceptual metrics (LPIPS, MUSIQ, NIQE, PI).

6. Comparison with SUPIR and FaithDiff at $\times 4$

We compare CASR with recent diffusion-based SR methods SUPIR [13] and FaithDiff [3] at the fixed $\times 4$ scale on DIV8K. As shown in Table 5, CASR achieves the best NIQE and PI scores, indicating strong perceptual naturalness. FaithDiff leads on PSNR, SSIM, LPIPS, and MUSIQ, which is expected as it is specifically optimized for fixed-scale reconstruction. In contrast, CASR targets arbitrary-

* Corresponding author.

Table 2. Comparison with ASISR methods on real-world datasets, with the best results in **bold**. Our approach archives consistently superior performance over others, showcasing strong generalization in real-world image.

Method	Benchmark														
	×8			×12			×18			×24			×30		
	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓
LINF [12]	37.25	9.22	8.13	27.29	10.17	8.93	21.68	11.41	9.82	20.13	12.50	10.48	19.11	13.34	10.95
BFSR[11]	37.25	9.22	8.13	27.29	10.17	8.93	21.68	11.41	9.82	20.13	12.50	10.48	19.11	13.34	10.95
IDM [6]	38.59	8.40	7.23	34.78	8.13	7.26	35.24	7.74	7.02	35.01	8.07	7.28	29.07	7.08	7.36
Kim [8]	36.61	8.79	7.89	35.15	9.22	8.58	31.42	9.81	9.47	30.11	10.09	9.63	27.08	10.03	9.74
LIIF [4] + Diff	40.99	7.92	7.38	29.90	9.28	8.37	23.44	10.81	9.30	21.02	11.71	9.87	19.60	12.40	10.28
CiaoSR [2] + Diff	41.97	8.04	7.47	31.49	9.21	8.33	24.53	10.39	9.04	21.71	11.36	9.64	19.73	11.96	9.99
CASR	64.12	5.56	4.71	59.72	5.42	4.63	55.56	6.09	5.15	50.85	6.47	5.55	47.53	6.74	5.76

Method	RealSRSet														
	×8			×12			×18			×24			×30		
	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓
LINF [12]	30.77	9.20	8.13	26.53	10.51	9.02	23.57	11.60	9.87	21.20	12.38	10.41	19.99	13.03	10.75
BFSR[11]	30.77	9.20	8.13	26.53	10.51	9.02	23.57	11.60	9.87	20.07	14.40	10.02	19.99	15.52	10.73
IDM [6]	31.39	7.82	6.88	29.13	8.10	7.11	28.15	7.88	6.90	31.33	8.29	7.40	29.66	7.17	7.38
Kim [8]	32.90	8.26	7.49	25.38	8.69	8.08	27.46	9.38	8.85	27.23	9.67	9.19	28.47	9.59	9.33
LIIF [4] + Diff	33.00	8.29	7.57	28.18	9.56	8.45	25.18	10.64	9.26	22.29	11.85	9.88	20.78	12.42	10.27
CiaoSR [2] + Diff	35.33	8.30	7.56	29.32	9.44	8.40	25.66	10.62	9.18	23.28	11.46	9.70	20.79	12.12	10.12
CASR	57.52	5.62	4.76	54.43	5.61	4.85	49.07	6.16	5.27	45.70	6.54	5.73	43.51	6.61	5.81

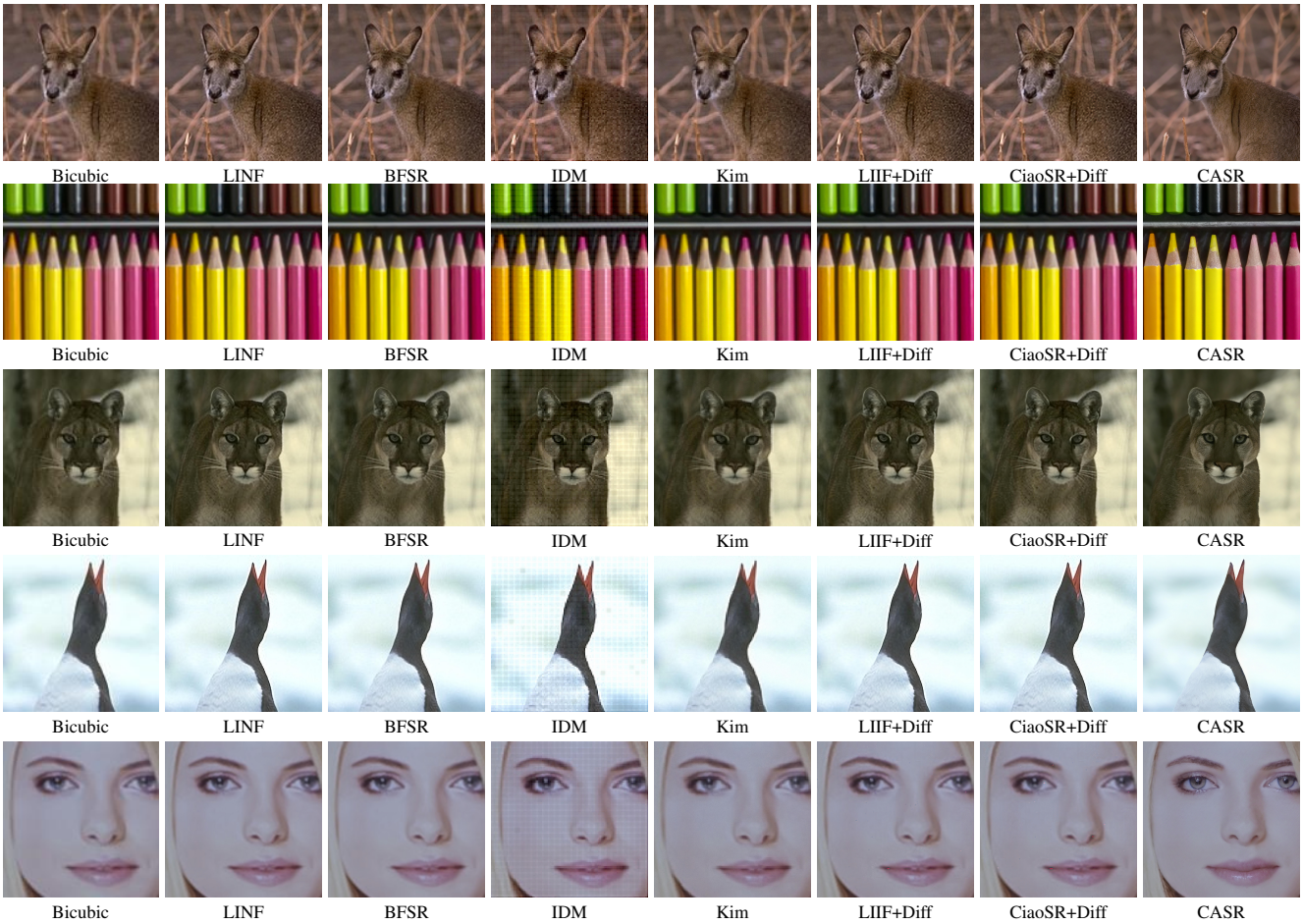

Figure 1. Qualitative comparison with different methods at ×24 magnification on the real-world dataset. Our method produces clearer and more natural results.

Table 3. Comparison on the COZ benchmark [5] using no-reference metrics. Best results in **bold**.

Method	×24			×30		
	MUSIQ↑	NIQE↓	PI↓	MUSIQ↑	NIQE↓	PI↓
Kim [8]	34.42	7.72	8.75	31.47	7.67	8.80
IDM [6]	38.83	8.83	7.93	37.66	8.24	8.09
CASR	**42.15**	**7.12**	**6.55**	**41.88**	**7.30**	**6.80**

Table 4. PSNR (dB) / SSIM on DIV8K across scales. CASR achieves the best fidelity among generative ASISR methods (IDM, Kim) at ×12 and above. INR-based and hybrid methods attain higher PSNR/SSIM via pixel-wise regression at the cost of perceptual quality (see Table 1 in the main paper).

Method	×8		×12		×18		×24		×30	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LINF [12]	28.83	.760	27.06	.724	25.64	.705	24.66	.695	23.90	.690
BFSR [11]	28.67	.752	27.00	.719	25.72	.702	24.78	.695	24.04	.690
LIIF [4]+Diff	28.49	.757	26.90	.723	25.66	.704	24.75	.695	23.95	.689
CiaoSR [2]+Diff	28.56	.760	26.87	.724	25.63	.705	24.64	.694	23.86	.689
IDM [6]	27.90	.724	22.23	.632	21.87	.626	21.17	.619	20.26	.623
Kim [8]	26.03	.657	21.14	.614	20.52	.598	20.71	.591	19.89	.639
CASR (Ours)	25.17	.676	24.16	.662	23.13	.657	22.57	.658	22.02	.653

Table 5. Comparison with diffusion-based SR methods at fixed ×4 scale on DIV8K. Best results in **bold**.

Method	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	NIQE↓	PI↓
SUPIR [13]	28.41	0.687	0.291	58.54	14.93	10.26
FaithDiff [3]	**28.68**	**0.711**	**0.244**	**61.14**	12.16	9.95
CASR	28.23	0.679	0.281	57.36	**12.03**	**9.72**

scale SR and is not tailored to any single scale, yet remains competitive overall.

7. Scale Decomposition Analysis

Table 6. Ablation study on scale decomposition strategies and their order.

Decomposition	DIV8K (×18)			
	LPIPS↓	MUSIQ↑	NIQE↓	PI↓
×18 (single step)	0.595	16.29	13.83	11.12
$\times 2 \times 2 \times 2 \times 1.5 \times 1.5$	0.532	48.32	7.14	6.97
$\times 2 \times 2 \times 3 \times 1.5$	0.512	50.97	7.47	6.45
$\times 1.5 \times 3 \times 4$	0.469	51.37	6.53	5.57
$\times 4 \times 3 \times 1.5$ (Ours)	**0.450**	**51.44**	**6.01**	**5.24**

We investigate how different decomposition strategies for the same overall factor (×18) affect performance. As shown in Table 6, single-step ×18 upsampling performs worst due to extreme out-of-distribution extrapolation. Among multi-step strategies, the *fewest stages, descending order* configuration ($\times 4 \times 3 \times 1.5$) achieves the best results, as larger initial steps exploit the model's full capacity while smaller later steps refine details with minimal error accumulation.

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2012. 1

[2] Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In *CVPR*, pages 1796–1807, 2023. 2, 3

[3] Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, pages 28188–28197, 2025. 1, 3

[4] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 2, 3

[5] Huiyuan Fu, Fei Peng, Xianwei Li, Yejun Li, Xin Wang, and Huadong Ma. Continuous optical zooming: A benchmark for arbitrary-scale image super-resolution in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3035–3044, 2024. 1, 3

[6] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *CVPR*, pages 10021–10030, 2023. 2, 3

[7] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 1

[8] Jinseok Kim and Tae-Kyun Kim. Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder. In *CVPR*, pages 9202–9211, 2024. 2, 3

[9] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *CVPR*, pages 10651–10662, 2022. 1

[10] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 416–423. IEEE, 2001. 1

[11] Li-Yuan Tsao, Yi-Chen Lo, Chia-Che Chang, Hao-Wei Chen, Roy Tseng, Chien Feng, and Chun-Yi Lee. Boosting flow-based generative super-resolution models via learned prior. In *CVPR*, pages 26005–26015, 2024. 2, 3

[12] Jie-En Yao, Li-Yuan Tsao, Yi-Chen Lo, Roy Tseng, Chia-Che Chang, and Chun-Yi Lee. Local implicit normalizing flow for arbitrary-scale image super-resolution. In *CVPR*, pages 1776–1785, 2023. 2, 3

[13] 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*, pages 25669–25680, 2024. 1, 3

[14] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 1

[15] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 1