Towards Photorealistic and Efficient Bokeh Rendering via Diffusion Framework

Supplementary Material

1. Degradation-aware Depth Estimation

1.1. Training Details

Despite the remarkable performance in HQ data, the accuracy of the depth estimation model deteriorates rapidly when applied to LQ images. And the input of imperfect disparity map degrades the results of SR bokeh rendering.

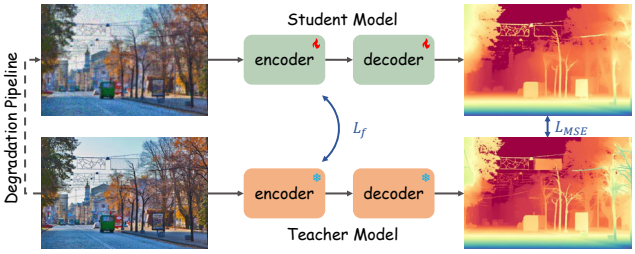

Figure 1. The training pipeline of the degradation-aware depth module.

To address this issue, we propose a self-feature distillation framework to estimate HQ-like features. As shown in Fig. 1, we utilize the pre-trained Depth Anything v2 Large model as the baseline network for both the teacher and the student models. During the training process, both HQ images and simulated degraded images are respectively input into the teacher and student networks to extract features from encoder. Through feature distillation, features are expected to remain consistent, thereby improving depth estimation performance. Simultaneously, the network's output is supervised to obtain a more accurate depth map.

1.2. Quantitative comparison of depth estimation

In our experiments, we use the pre-trained Depth Anything v2 as the teacher model to generate pseudo-labels and supervise the student model, initialized identically, within a distillation framework that takes only RGB images as input. Specifically, we conduct our distillation experiments using a subset of 200,000 samples from the SA-1B dataset [3]. The Real-ESRGAN degradation pipeline [9] is used to synthesize LQ-HQ training pairs.

To demonstrate the effectiveness of our degradation-aware depth model on degraded images, we compare our approach with the baseline method, Depth Anything v2. Tab. 1 shows that our method outperforms these related works on the degraded NYUv2 [8] (for indoor scenes) and KITTI [2] (for outdoor scenes). We use point prompts for "Degrade", "Clear" and "Average" scenarios. "Degrade" refers to images degraded by Real-ESRGAN, "Clear" refers to the original, non-degraded images, and "Average" is the mean value of the "Degrade" and "Clear" images. The bold values indicate the best performance. Through self-feature distillation, our student model not only exhibits minimal performance degradation on clear images but also outperforms the baseline on degraded images, thereby verifying the superiority of our method.

2. Detail of bokeh training datasets

To obtain HQ bokeh images as ground truth in bokeh training stage, similar to MPIB [6] and Dr.Bokeh [7], we built a ray-tracing-based renderer that generates lens blur through a real thin lens, as shown in Fig. 2. We first collected nearly 2k high-resolution landscape images from the Internet to serve as our background images. The foreground images are collected from PhotoMatte85 [5], RWP-636 [11], AIM-500 [4] and websites. Each sample is randomly composed of two selected foreground images and one background image. During the composition process, the disparity map is set within the range from 0 to 1, the random blur parameter ranges from 0 to 32, and the disparity focus is randomly set to one of the positions in either the foreground or the background. In order to introduce more variation in depth and create more diverse blur effects in the training data, we randomly set the depth variation for the background.

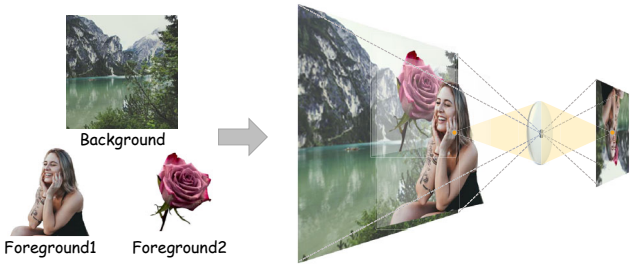

Figure 2. The pipeline of data synthesis.

3. Quantitative comparison on Real-ISR

Although our method is not specifically designed for super-resolution tasks, setting the blur intensity K to 0 allows us to obtain all-in-focus HR images. Furthermore, despite not incorporating text conditions, MagicBokeh still shows performance in the single task of Real-ISR, as illustrated in Tab. 2, highlighting its ability to restore both the realism and aesthetic quality of images.

4. More Results

4.1. Adjusting Aperture

We present the results of increased blurriness in Fig. 3. MagicBokeh successfully achieves progressive blurriness

Table 1. Quantitative comparison on the NYUv2 and KITTI datasets (seen datasets with synthetic degradations) for "Degrade", "Clear", and "Average" scenarios.

Dataset	Method	Degrade		Clear		Average	
		AbsRel ↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
NYUv2	Depth Anything v2	0.081	0.926	**0.043**	**0.981**	0.062	0.954
	DA depth	**0.068**	**0.946**	0.047	0.976	**0.058**	**0.961**
KITTI	Depth Anything v2	0.123	0.852	**0.074**	**0.946**	0.099	0.899
	DA depth	**0.105**	**0.883**	0.079	0.944	**0.092**	**0.914**

Table 2. Quantitative comparison with state-of-the-art methods on real-world benchmarks (RealSR [1] and DrealSR [10]). By providing a defocus map with all-zero input, our method can generate a high-quality all-in-focus image for quantitative comparison. The best and second-best results are highlighted in **red** and blue.

Datasets	Methods	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	CLIP-IQA ↑	NIQE ↓	MUSIQ ↑	MANIQA ↑	FID ↓
RealSR	SinSR	**26.32**	0.7363	0.3195	0.2351	0.6153	6.3541	60.42	0.5366	138.64
	OSEDiff	25.15	0.7341	0.2921	0.2128	0.6685	5.6528	**69.11**	0.6332	123.68
	S3Diff	25.18	0.7269	**0.2722**	**0.2005**	**0.6742**	**5.2612**	67.82	**0.6417**	**105.11**
	MagicBokeh	26.14	**0.7392**	0.2888	0.2192	0.6246	5.6337	67.22	0.6214	123.06
DRealSR	SinSR	28.35	0.7484	0.3689	0.2497	0.6319	6.9533	55.09	0.4881	170.18
	OSEDiff	27.91	0.7834	**0.2968**	0.2269	0.6964	6.4907	**64.65**	0.5899	135.28
	S3Diff	27.54	0.7491	0.3109	**0.2100**	**0.7132**	6.1935	63.93	**0.6099**	**118.57**
	MagicBokeh	**28.99**	**0.7901**	0.3003	0.2220	0.6633	**6.1204**	62.93	0.5901	143.02

while maintaining subject sharpness. The cases are high-zoom real mobile device captures, and MagicBokeh generates realistic bokeh effects.

4.2. Adjusting Focus Distance

We provide examples of changing focus distance in Fig. 4. Whether focusing on the foreground or background, our method can achieve natural super-resolution and bokeh effects.

4.3. More Comparisons

Here, we provide more comparisons between MagicBokeh and other two-stage pipeline to further validate the effectiveness of MagicBokeh. First, we demonstrate more comparisons in Fig. 5. In the first example, MagicBokeh produces bokeh effects that are closer to the Ground Truth compared to other methods, especially in the red-boxed area. Compared to methods including BokehMe and Dr.Bokeh in the green-boxed area, our method and BokehDiff generate sharper edges. In the second example, in terms of super-resolution, our method produces more distinct leaf details compared to OSEDiff and S3Diff. In terms of bokeh, our method generates the best edge effects compared to BokehDiff, BokehMe, and Dr.Bokeh. In the third example, our method can still produce bokeh effects that are consistent with the real situation, even in the presence of noise. We continue the demonstration of results in Fig. 6. Our method gradually increases the blur with increasing defocus while keeping the focused foreground unchanged, resulting in a more realistic effect.

References

[1] 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*, pages 3086–3095, 2019. 2

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1

[4] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 1

[5] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 1

[6] Juewen Peng, Jianming Zhang, Xianrui Luo, Hao Lu, Ke Xian, and Zhiguo Cao. Mpib: An mpi-based bokeh rendering framework for realistic partial occlusion effects. In

Figure 3. Given the defocus map and LR input, our method is able to gradually increase the aperture parameter from 1x blur to 3x blur (*Zoom-in for best view*).

Figure 4. Given the disparity map and LR input, our method is able to achieve dynamic adjustment of the focus distance (*Zoom-in for best view*).

European Conference on Computer Vision, pages 590–607. Springer, 2022. 1

[7] Yichen Sheng, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin, and Bedrich Benes. Dr. bokeh: differentiable occlusion-aware bokeh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4515–4525, 2024. 1

[8] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 1

[9] 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. 1

[10] 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

[11] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1154–1163, 2021. 1

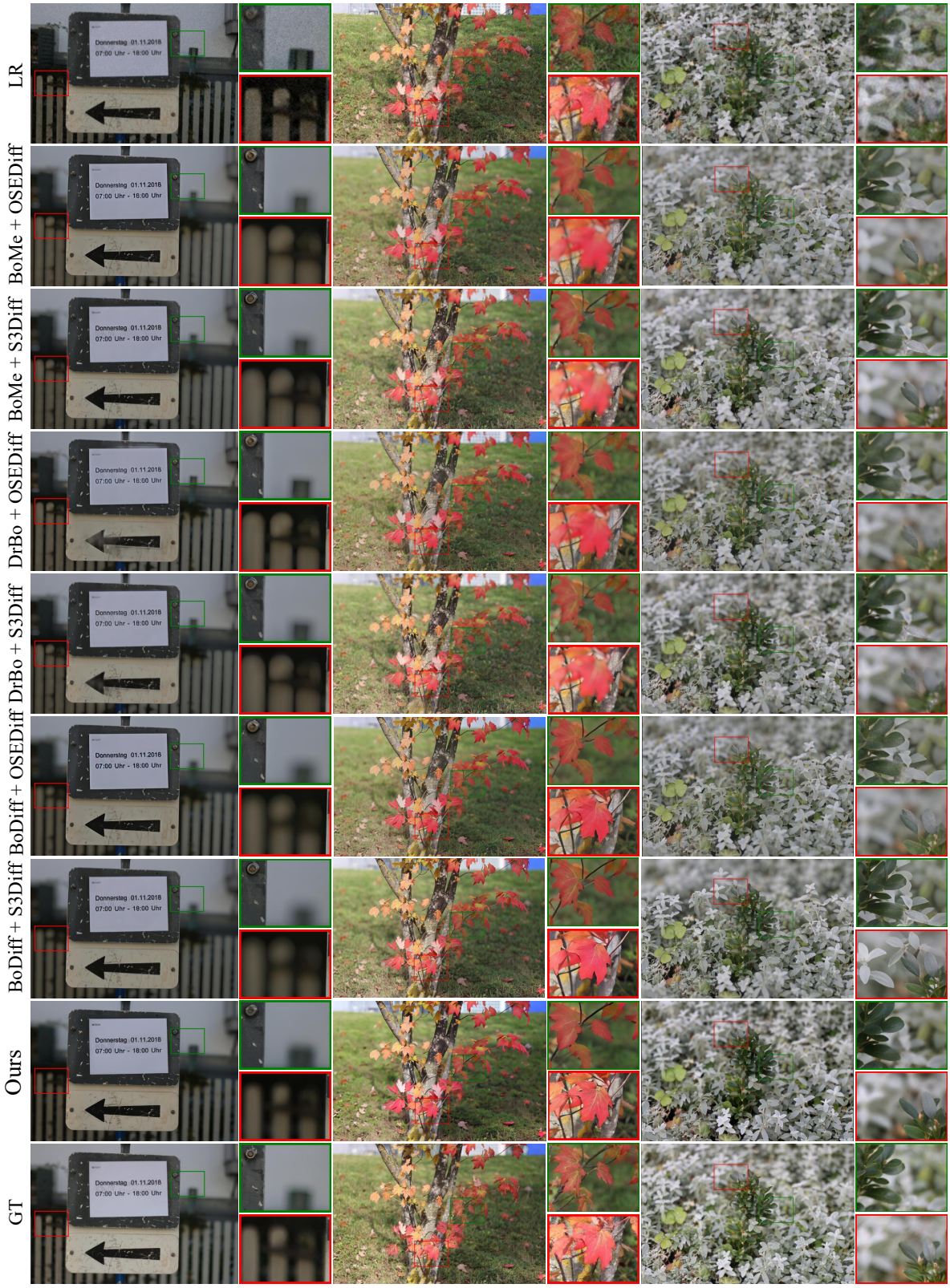

Figure 5. Qualitative comparison on EBB400-LQ (*Zoom-in for best view*).

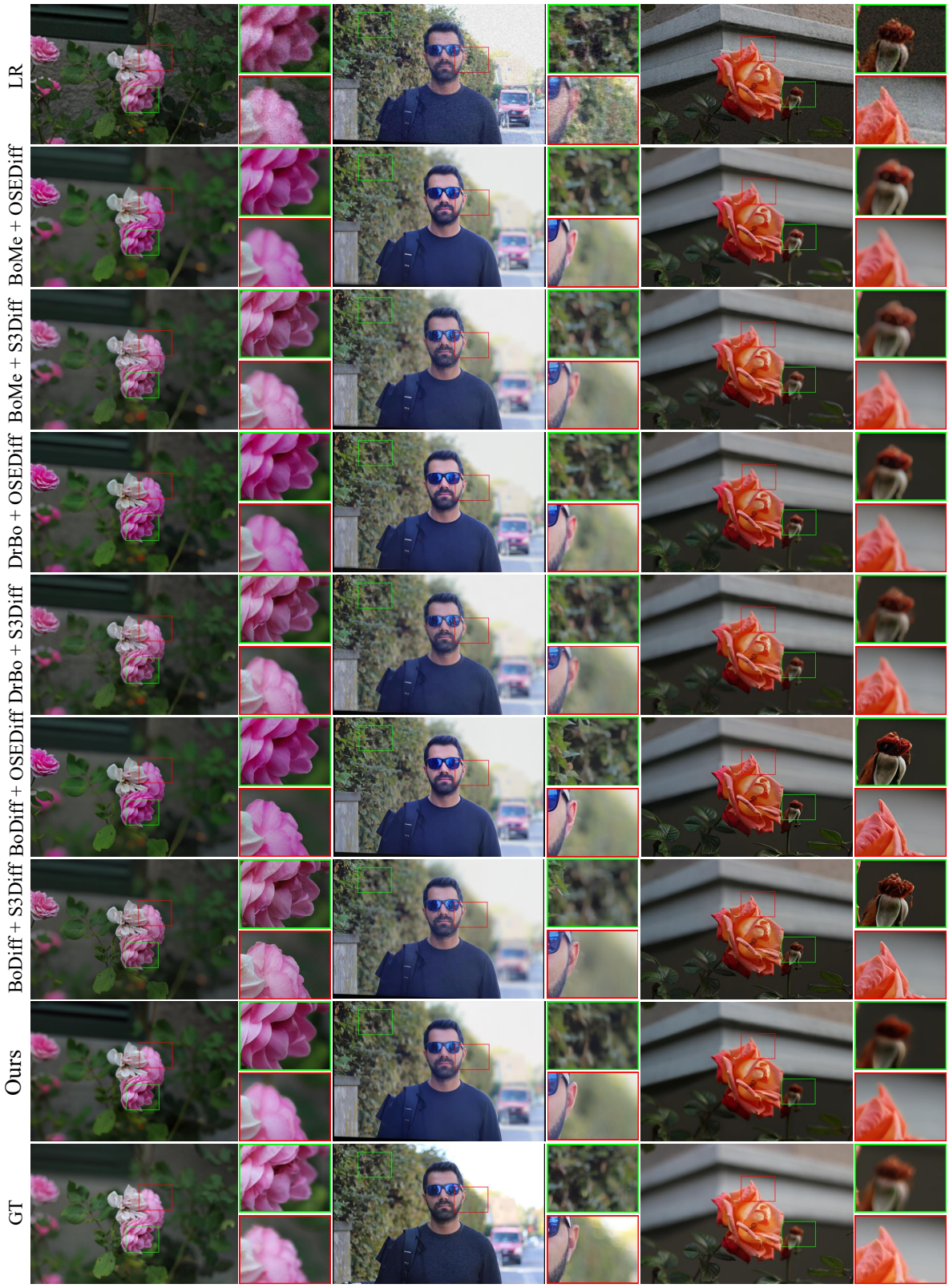

Figure 6. Qualitative comparison on EBB400-LQ (*Zoom-in for best view*).