Fast Image Super-Resolution via Consistency Rectified Flow
Supplementary Material

In this supplementary material, we first provide additional details about our FlowSR in Sec. 1. Next, we present more experimental results in Sec. 2. Finally, we discuss the limitations of our approach and outline potential future directions in Sec. 3.

1. Implementation Details

We first fine-tune the pre-trained SD model [6] to adapt it to our SR flow learning objectives. The fine-tuned SR flow model is then used to initialize both the SR model θ and the teacher model ϕ. A default text prompt is used for the SD model. During consistency SR flow training, each training batch is split into two groups: one for SR flow learning and the other for consistency learning. This approach ensures that the fine-tuned SR model still learns accurate SR flow while also acquiring distilled one-step high-quality inference capability.

For the fast-slow time scheduling, the adjacent time steps t and $t' = t + \Delta t$ are sampled as follows: we first randomly select either the fast scheduler or the slow scheduler and use it to sample t'. Then, the other scheduler is used to sample t. If the fast scheduler is chosen first, t is sampled from the range between t' and its predecessor timestep. Conversely, if the slow scheduler is chosen first, t is sampled from the next time point less than t'. This approach ensures that the jump Δt remains flexible.

We also observe that the choice of timestep shifting and sampling plays a crucial role in SR flow learning, and we provide an ablation study in Sec. 2.4 to further analyze this.

For the image quality alignment loss, we employ Qwen2-VL [9] to generate image quality captions. Alternatively, other MLLMs or fixed quality description prompts can also be used to compute this loss.

2. More Results

2.1. Evaluation on DIV2K-Val

We also evaluate our method on the DIV2K-Val dataset [1, 8]. Table 1 provides a quantitative comparison of various SR methods. Across all reference-based metrics, our FlowSR achieves state-of-the-art performance or performs on par with the best existing methods. For no-reference metrics, while FlowSR performs worse than the multi-step SD-based PASD [14], it remains the best-performing model among all single-step sampling methods. These results demonstrate the effectiveness and superiority of our method.

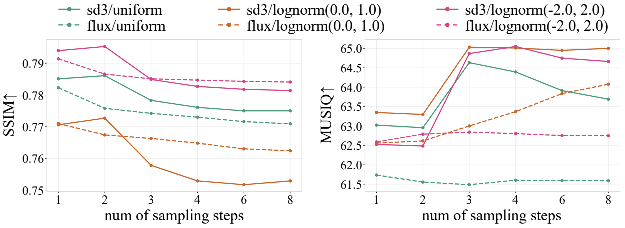

Figure 1. Impact of timestep shifting / timestep sampling. SD3 timestep shifting with `lognorm(-2.0, 2.0)` timestep sampling achieves a good fidelity/quality tradeoff on DRealSR [11].

2.2. Model efficiency

We present the model parameters, MACs, and latency in Table 2. The MACs and runtime are measured for 4× SR using a 128×128 LR input. Note that we use a fixed text prompt for model inference, eliminating the need for text encoding in the SD model. As demonstrated, our method shows a significant advantage over multi-step SR approaches, such as StableSR and SeeSR, while maintaining comparable computational complexity to one-step methods like OSEDiff.

2.3. More Qualitative Visual Comparisons

Figs. 2 to 4 provide additional visual comparisons between FlowSR and other DM-based SR methods. Our visual results are consistently better than, or at least comparable to, all multi-step and single-step diffusion methods across various scenarios, such as flowers, buildings, and clothing. Visual comparisons also support the conclusions drawn from the quantitative study, highlighting the higher fidelity of our results. Overall, FlowSR exhibits more natural details, along with realistic textures and structures.

Table 1. Quantitative comparisons of different methods on the DIV2K-Val dataset.

Methods	#Steps	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	CLIPIQA ↑
StableSR [8]	200	23.26	0.5726	0.3113	0.2048	**24.44**	4.76	65.92	0.6192	0.6771
DiffBIR [5]	50	23.64	0.5647	0.3524	0.2128	30.72	4.70	65.81	0.6210	0.6704
SeeSR [13]	50	23.68	0.6043	0.3194	0.1968	25.90	4.81	68.67	0.6240	**0.6936**
PASD [14]	20	23.14	0.5505	0.3571	0.2207	29.20	**4.36**	**68.95**	**0.6483**	0.6788
ResShift [15]	15	**24.65**	0.6181	0.3349	0.2213	36.11	6.82	61.09	0.5454	0.6071
SinSR [10]	1	24.41	0.6018	0.3240	0.2066	35.57	6.02	62.82	0.5386	0.6471
OSEDiff [12]	1	23.72	0.6108	0.2941	0.1976	26.32	4.71	67.97	0.6148	0.6683
DoSSR [2]	1	24.35	**0.6265**	0.3725	0.2786	50.27	10.38	58.44	0.5024	0.6187
FlowSR	1	24.42	0.6192	**0.2798**	**0.1847**	24.52	4.63	68.22	0.6193	0.6901

Table 2. Efficiency metrics of parameters, MACs, and runtime.

Method	StableSR	DiffBIR	SeeSR	PASD	ResShift	SinSR	OSEDiff	DoSSR	FlowSR
#steps ↓	200	50	50	20	15	1	1	1	1
#param (M) ↓	1409	1683	2511	2314	174	174	1765	1718	982
MACs (G) ↓	95382	24234	66444	23592	4962	2119	2323	3232	2148
time (s) ↓	13.54	6.51	5.21	3.47	0.89	0.13	0.16	0.28	0.14

2.4. Impact of timestep shifting and sampling

We train the basic SR flow models using different time scheduling methods to evaluate their impact. We select representative timestep shifting options, including SD3 [3], which biases timesteps toward $t = 1$, and FLUX.1-schnell[1], which uses uniform timesteps. For timestep sampling, we use `lognorm(0.0, 1.0)` as adopted in [3], `lognorm(-2.0, 2.0)` studied in [7], and uniform sampling. The first sampling method favors intermediate timesteps, while the second samples more timesteps closer to $t = 1$. The results for different inference steps are shown in Fig. 1. We observe that: (1) SD3 timesteps outperform the uniform timesteps for SR flow in most cases; (2) `lognorm(0.0, 1.0)` achieves high quality (MUSIQ) but sacrifices fidelity (SSIM). In our experiments, we employ SD3 timesteps with `lognorm(-2.0, 2.0)` timestep sampling, as it demonstrates high fidelity with one-step inference and good quality with few-step inference.

3. Limitations and Future Works

In this work, we tackle one-step SR from the perspective of flow and consistency. We provide valuable insights into the effective use of flow-based techniques and consistency learning to achieve competitive SR results in a single-step setting. While our study demonstrates promising results, there are some limitations. First, due to computational constraints, we have not yet explored more advanced T2I models, such as SD3 [3] and FLUX [4], as potential backbones. Second, we are actively working on further reducing the number of parameters in the backbone network to achieve additional efficiency gains.

[1] https://huggingface.co/black-forest-labs/FLUX.1-schnell

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 1

[2] Qinpeng Cui, Yixuan Liu, Xinyi Zhang, Qiqi Bao, Zhongdao Wang, Qingmin Liao, Li Wang, Tian Lu, and Emad Barsoum. Taming diffusion prior for image super-resolution with domain shift sdes. *NeurIPS*, 2024. 2

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2

[4] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[5] 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. 2

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[7] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia*, 2024. 2

[8] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024. 1, 2

[9] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[10] 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. 2

[11] 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

[12] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *NeurIPS*, 2024. 2

[13] 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. 2

[14] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, 2024. 1, 2

[15] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 2023. 2

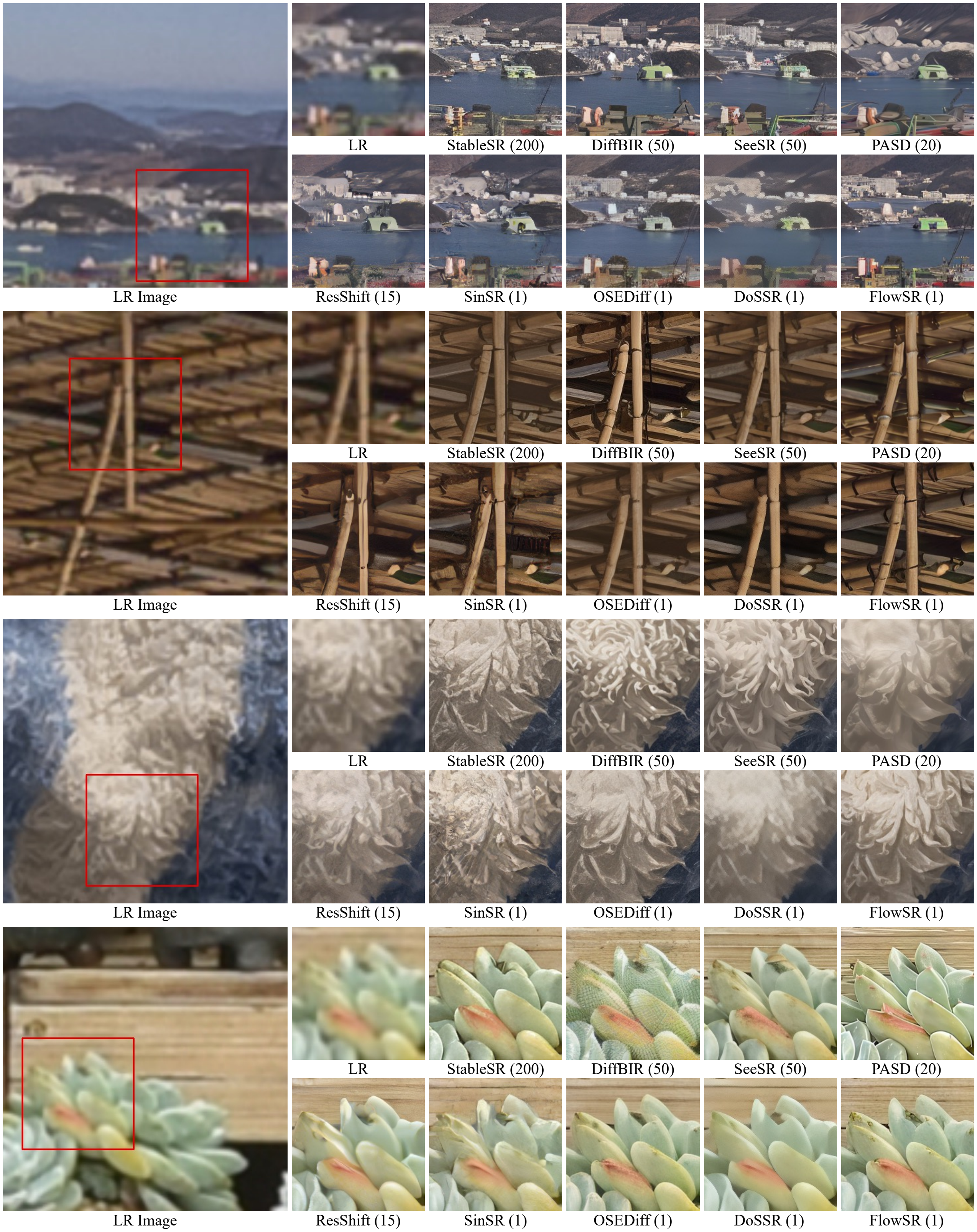

Figure 2. Visual comparisons of different SR methods on real-world examples #1. The number of sampling steps are indicated in bracket. Please zoom in for a better view.

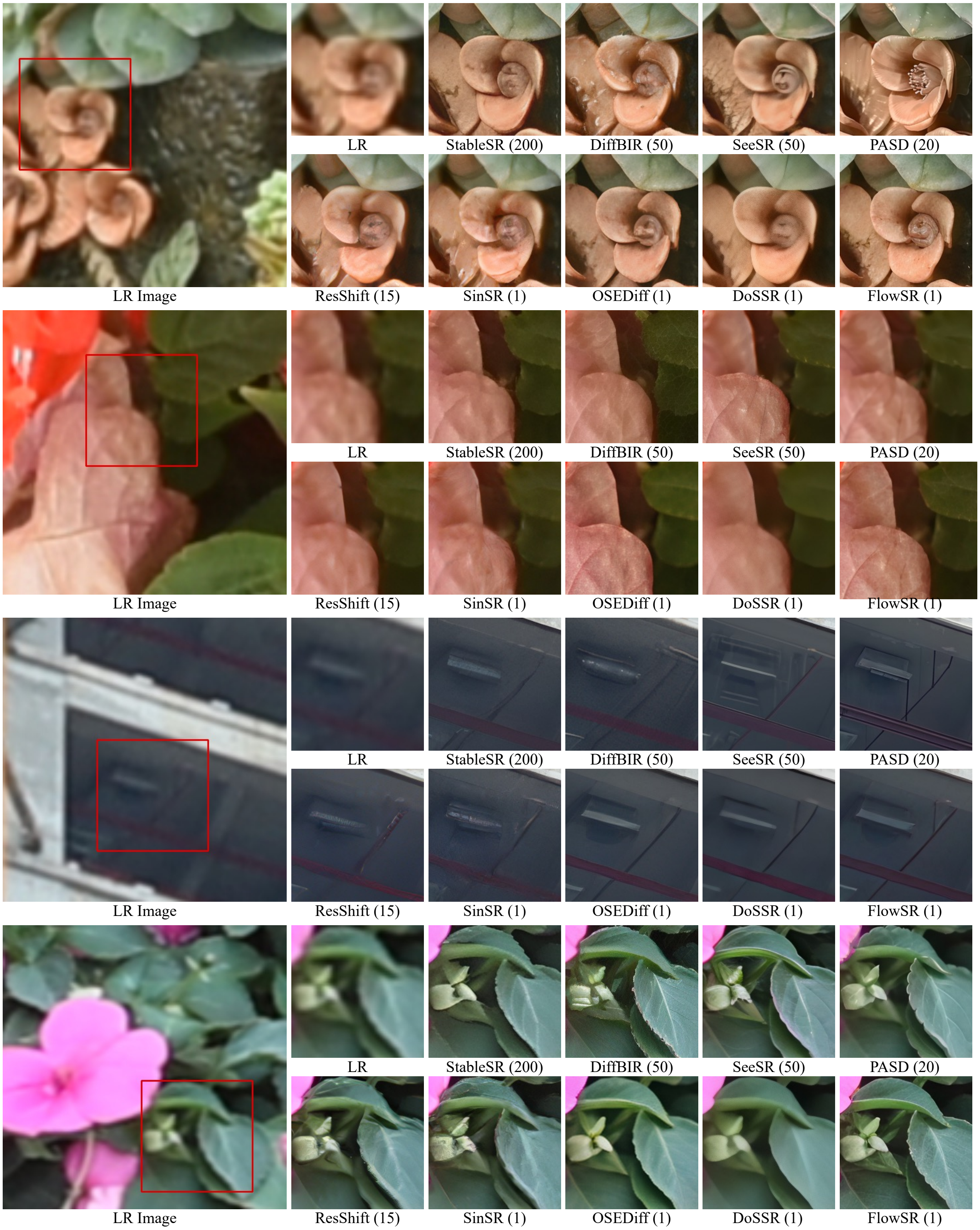

Figure 3. Visual comparisons of different SR methods on real-world examples #2. The number of sampling steps are indicated in bracket. Please zoom in for a better view.

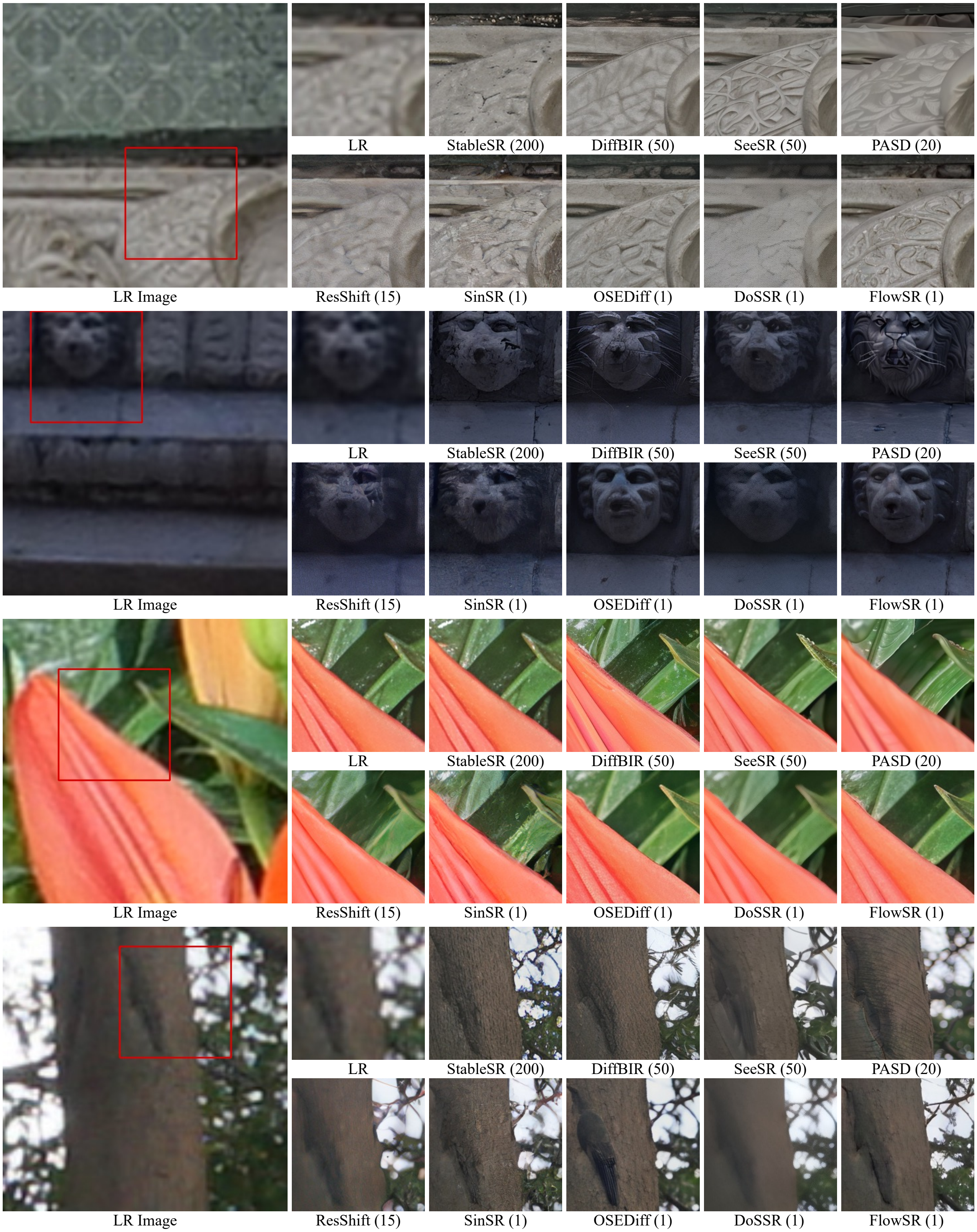

Figure 4. Visual comparisons of different SR methods on real-world examples #3. The number of sampling steps are indicated in bracket. Please zoom in for a better view.