

A Supplementary Materials

In the supplementary materials, we provide the following materials:

1. Ablation experiments on pruning locations.
2. The effectiveness of combining channel pruning with multi-layer feature distillation and the selection of pruning ratios.
3. Performance on the DIV2K validation set [1].
4. Comparisons with non-diffusion SR models.
5. Performance on high-resolution SR.
6. The effectiveness of Stable Diffusion prior.
7. Detailed ablation studies for the Adaptive Skip Connection (ASC).
8. Explanation and ablation studies for the Dual-path Feature Injection (DFI).
9. Additional quantitative comparisons.
10. Details of the lite decoder structure.
11. Details of our training loss.
12. More qualitative comparisons.

A.1 Ablation Experiments on Pruning Locations

As discussed in Section 3.2.2, the choice of pruning location plays a critical role in model performance. In particular, pruning shallower modules has a greater impact on super-resolution quality, whereas deeper blocks can be pruned with relatively minor performance degradation. We conduct ablation studies on four computationally intensive components within the diffusion U-Net: residual blocks, cross-attention layers, self-attention layers, and feed-forward networks (FFNs). U-Net depth levels are categorized as I, II, III, IV, from shallowest to deepest. All experiments are performed on the RealSR dataset [2].

Table 1: Comparison of quantitative performance and efficiency across residual block pruning locations. Our selected setting is marked in **bold**.

Pruning Position	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	FID \downarrow	Time (ms)	MACs (G)	Param. (M)
None	24.84	0.7232	0.2653	0.2015	110.46	17.73	279	430
Depth IV	25.08	0.7307	0.2673	0.2047	116.76	17.32	278	422
Depth III, IV (Ours)	24.98	0.7294	0.2695	0.2041	117.76	15.79	268	382
Depth II, III, IV	24.51	0.7060	0.2866	0.2162	133.09	14.16	258	372
Depth I, II, III, IV	24.25	0.6929	0.3019	0.2253	131.81	12.59	250	370

Table 2: Comparison of quantitative performance and efficiency across cross-attention layer pruning locations. Our selected setting is marked in **bold**.

Pruning Position	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	FID \downarrow	Time (ms)	MACs (G)	Param. (M)
None	24.84	0.7232	0.2653	0.2015	110.46	17.73	279	430
Depth IV	25.14	0.7313	0.2676	0.2052	116.21	17.33	278	427
Depth III, IV	24.92	0.7270	0.2671	0.2032	112.89	16.71	276	410
Depth II, III, IV	24.76	0.7224	0.2682	0.2026	111.07	15.91	273	404
Depth I, II, III, IV (Ours)	24.97	0.7307	0.2666	0.2031	112.85	14.97	271	401

As shown in Table 1, pruning residual blocks at shallow depths (I and II) results in significant degradation in perceptual metrics such as LPIPS, DISTS, and FID. Consequently, we limit pruning to deeper layers (depths III and IV). The ablation study on cross-attention layers, presented in Table 2, reveals that their contribution is limited, as super-resolution primarily depends on fine-grained visual

details rather than high-level semantic alignment. We find that pruning all cross-attention layers has minimal impact on overall model quality.

Table 3 reports the ablation results for self-attention layer pruning. Pruning at depths III and above leads to a noticeable decline in perceptual metrics. While PSNR and SSIM slightly improve when pruning layers at depths III and IV, we attribute this to oversmoothing effects rather than actual perceptual gains. Lastly, we examine the pruning positions of feed-forward networks (Table 4). Balancing performance and efficiency, we opt to prune FFNs at depths III and IV.

Table 3: Comparison of quantitative performance and efficiency across self-attention layer pruning locations. Our selected setting is marked in **bold**.

Pruning Position	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	Time (ms)	MACs (G)	Param. (M)
None	24.84	0.7232	0.2653	0.2015	110.46	17.73	279	430
Depth IV (Ours)	25.14	0.7315	0.2677	0.2056	116.56	17.49	279	430
Depth III, IV	25.43	0.7308	0.2757	0.2257	147.15	17.08	278	430
Depth II, III, IV	23.51	0.6428	0.3583	0.2640	196.66	16.54	274	430
Depth I, II, III, IV	22.50	0.5310	0.4611	0.3038	244.87	13.18	242	430

Table 4: Comparison of quantitative performance and efficiency across feed-forward network pruning locations. Our selected setting is marked in **bold**.

Pruning Position	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	Time (ms)	MACs (G)	Param. (M)
None	24.84	0.7232	0.2653	0.2015	110.46	17.73	279	430
Depth IV	25.08	0.7307	0.2673	0.2047	116.76	17.32	278	422
Depth III, IV (Ours)	24.98	0.7294	0.2695	0.2041	117.76	15.79	268	382
Depth II, III, IV	24.51	0.7060	0.2866	0.2162	133.09	14.16	258	372
Depth I, II, III, IV	24.25	0.6929	0.3019	0.2253	131.81	12.59	250	370

A.2 Channel Pruning with Multi-layer Feature Distillation

Table 5 presents the impact of the proposed multi-layer feature distillation on channel pruning, alongside performance and efficiency comparisons across varying pruning ratios. As the pruning ratio increases, model efficiency improves, but perceptual quality gradually declines. Figure 1 further illustrates that pruning without multi-layer feature distillation leads to a significant performance drop, whereas incorporating it effectively mitigates this degradation by preserving prior knowledge that would otherwise be lost. Balancing performance and efficiency, we adopt 30% channel pruning with multi-layer feature distillation as our final configuration.

Table 5: Comparison of performance and efficiency of channel pruning at different ratios with and without multi-layer feature distillation. The selected setting is highlighted in **bold**.

Pruning Ratio	w/ Multi-layer Distillation			w/o Multi-layer Distillation			Time (ms)	MACs (G)	Param. (M)
	LPIPS↓	DISTS↓	FID↓	LPIPS↓	DISTS↓	FID↓			
0%	0.2474	0.1911	101.82	0.2474	0.1911	101.82	31.37	482	868
10%	0.2625	0.1983	103.80	0.2669	0.2048	106.95	25.78	399	704
20%	0.2647	0.1981	107.11	0.2618	0.1983	115.21	22.19	347	561
30%	0.2653	0.2015	110.46	0.2653	0.2057	126.47	17.73	279	430
40%	0.2712	0.2058	116.51	0.2798	0.2148	140.85	15.25	239	320
50%	0.2865	0.2126	126.53	0.3090	0.2290	156.72	11.24	186	224

A.3 Performance on the DIV2K Validation Set

To further assess the robustness of our method under more diverse degradations, we evaluate it on the DIV2K validation set, with results presented in Table 6. This dataset, introduced by StableSR [1],

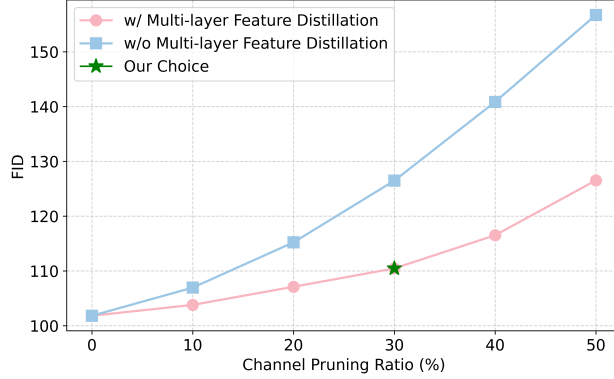


Figure 1: FID comparison of channel pruning at different ratios with and without Multi-layer Feature Distillation.

contains low-resolution images generated using Real-ESRGAN [3]. PocketSR surpasses previous single-step and even multi-step methods in LPIPS and DISTs, ranks second in FID and NIQE, and performs comparably to other single-step approaches on the remaining metrics. These results highlight PocketSR’s strong perceptual quality and its robustness to challenging degradations.

Table 6: Quantitative comparisons with state-of-the-art methods on the DIV2K validation set. The best results for each metric are highlighted in **bold**, and the second-best are underlined.

Datasets	Metrics	StableSR [1]	DiffBIR [4]	SeeSR [5]	ResShift [6]	SinSR [7]	OSDiff [8]	AdcSR [9]	PocketSR
DIV2K Val. [1]	LPIPS↓	0.3113	0.3524	0.3194	0.3349	0.3240	0.2941	<u>0.2853</u>	0.2801
	DISTS↓	0.2048	0.2128	0.1968	0.2213	0.2066	0.1976	<u>0.1899</u>	0.1830
	PSNR↑	23.26	23.64	23.68	24.65	<u>24.41</u>	23.72	23.74	23.85
	SSIM↑	0.5726	0.5647	0.6043	0.6181	0.6018	0.6108	0.6017	0.6015
	FID↓	24.44	30.72	25.90	36.11	35.57	26.32	25.52	<u>25.25</u>
	NIQE↓	4.760	4.700	4.810	6.820	6.020	4.710	4.360	<u>4.415</u>
	MUSIQ↑	65.92	65.81	68.67	61.09	62.82	67.97	<u>68.00</u>	<u>66.38</u>

A.4 Comparisons with non-diffusion SR models

We have conducted additional comparisons on the DRealSR dataset with several representative non-diffusion-based methods, including a CNN-based method (DASR [10]) and a Transformer-based method (FeMaSR [11]). The comparison results are presented in Table 7 (FPS measured on an A100 server).

Our method significantly outperforms the competing approaches in terms of perceptual quality metrics such as LPIPS, DISTs, NIQE, and MUSIQ. While DASR shows stronger results in PSNR, SSIM, and efficiency, we believe this is partly due to its extremely lightweight design, which—while advantageous in terms of speed and model size—also limits its ability to recover fine-grained details from complex real-world degradations, resulting in smoother outputs. Such characteristics can sometimes lead to higher PSNR and SSIM scores that may not fully reflect perceptual quality.

A.5 Performance on high-resolution SR

To assess PocketSR at higher resolutions, we constructed a multi-resolution benchmark with three 4× SR tasks: upscaling to 1K, 2K, and 4K. Each setting includes 20 test images with ground-truths. The 1K and 2K samples are selected from RealSR, ensuring resolution and scene diversity; 4K samples are from DRealSR. Notably, this benchmark uses original full-resolution images, unlike Table 2 in the main text, which evaluates on center-cropped patches—a common practice in SR [1].

As shown in Table 8, PocketSR consistently achieves superior LPIPS, DISTs, and NIQE scores, and competitive PSNR, SSIM, and MUSIQ results. Its stable performance across resolutions demonstrates strong robustness, making it well-suited for real-world use cases like mobile photography.

Table 7: Quantitative and efficiency comparisons with representative non-diffusion SR methods on DRealSR. The best results for each metric are highlighted in bold.

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	MUSIQ \uparrow	Params (M)	MACs (G)	FPS
DASR [10]	29.75	0.8262	0.3099	0.2275	7.586	42.41	8	46	148.0
FeMaSR [11]	26.87	0.7569	0.3157	0.2239	5.910	53.71	34	384	16.8
PocketSR (Ours)	28.05	0.7675	0.2962	0.2139	5.809	63.85	146	225	62.5

Table 8: Comparison of different super-resolution methods across various resolutions.

Tasks	Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	MUSIQ \uparrow
1K Resolution SR	SinSR	25.53	0.7015	0.3934	0.2286	5.794	62.82
	OSDiff	24.73	0.7088	0.3226	0.1933	3.943	70.50
	AdcSR	25.01	0.6995	0.3154	0.1886	3.409	72.05
	PocketSR (Ours)	25.07	0.7162	0.2709	0.1662	3.880	65.23
2K Resolution SR	SinSR	26.96	0.7473	0.3884	0.2102	6.563	55.25
	OSDiff	25.81	0.7756	0.2775	0.1661	4.333	63.78
	AdcSR	26.35	0.7732	0.2729	0.1630	4.158	63.52
	PocketSR (Ours)	26.41	0.7736	0.2467	0.1510	4.060	63.85
4K Resolution SR	SinSR	27.36	0.7086	0.4753	0.2379	6.246	29.56
	OSDiff	26.37	0.7951	0.3200	0.1696	4.291	36.69
	AdcSR	27.37	0.7888	0.3258	0.1784	4.325	34.34
	PocketSR (Ours)	26.76	0.7752	0.3174	0.1664	4.064	34.25

A.6 Effect of Stable Diffusion Prior

PocketSR is built on a pre-trained Stable Diffusion (SD) model, followed by pruning and distillation. To assess whether PocketSR successfully preserves the prior knowledge from the pre-trained SD, we conduct an ablation study on the SD prior, as shown in Table 9. Specifically, we retain the pruned architecture of PocketSR but remove the SD-based initialization for the U-Net, training the entire model from scratch instead. This variant is referred to as “w/o SD prior”. Results show that leveraging the SD prior yields superior performance across all metrics except NIQE, with a notable PSNR improvement of nearly 0.8 dB. The higher NIQE score is likely due to GAN-induced artifacts arising from the lack of SD prior guidance.

Table 9: Ablation Study of Stable Diffusion prior on the RealSR and DRealSR datasets. The best results for each metric are highlighted in bold.

Datasets	Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	MUSIQ \uparrow
RealSR	Ours	25.47	0.7330	0.2713	0.2094	5.067	67.07
	w/o SD Prior	24.66	0.7080	0.2718	0.2132	4.912	65.72
DRealSR	Ours	28.05	0.7675	0.2962	0.2139	5.809	63.85
	w/o SD Prior	27.22	0.7364	0.3034	0.2181	5.697	62.32

A.7 Detailed ablation studies for the Adaptive Skip Connection (ASC)

We provide additional ablation studies on the Adaptive Skip Connection (ASC) module in Table 10. The experiments are conducted on the RealSR dataset, following the same settings as in Table 3 of the main paper for ease of comparison. We omit efficiency-related metrics here, as the differences are minimal.

We investigate the effect of removing the learnable control coefficients in ASC. We observe that using naive skip connections hinders the model’s ability to adaptively control the contribution of skip features based on the degradation level of the input. As a result, the model performs poorly when handling diverse real-world degradations.

Table 10: Ablation studies for the Adaptive Skip Connection.

LiteED Design	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow
Original	25.61	0.7431	0.2474	0.1911	5.200
w/o control coefficients in ASC	25.16	0.7276	0.2486	0.1999	5.321
w/ control coefficients learning from the heavy path	25.08	0.7347	0.2667	0.1995	4.952

Table 11: Ablation studies for the Dual-path Feature Injection (DFI).

LiteED Design	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	Time (ms)	MACs (G)	Param. (M)
Original DFI	25.61	0.7431	0.2474	0.1911	5.200	31.4	481.6	868.3
w/o Lite Path	24.75	0.7229	0.2596	0.1991	4.860	31.3	481.6	868.3
w/o Heavy Path	25.35	0.7427	0.2580	0.1940	5.136	31.1	478.6	867.6

We explore an alternative design where the control coefficients are predicted from features in the heavy path. However, we find that this strategy leads to instability during training, as heavy path features are uncompressed and contain redundant information, which interferes with the learning of accurate control signals. Although this variant performs relatively well on the no-reference metric NIQE, the significantly lower LPIPS and PSNR scores suggest compromised perceptual and pixel-level fidelity, with more noticeable artifacts in the output.

A.8 Explanation and ablation studies for the Dual-path Feature Injection (DFI)

In our Dual-path Feature Injection (DFI) design, the lite path provides compressed, information-dense features that offer global structural guidance and align well with the VAE feature distribution in SD, making them easier for the pretrained U-Net to utilize. In contrast, the heavy path contains richer details but lower information density, making it harder for generative models to use directly.

The lite path serves as the primary guidance source, while the heavy path complements it with fine-grained textures. Both are essential for high-quality super-resolution. To validate this, we conducted ablation studies following the experimental settings in Table 3 of the main paper. The results are shown in Table 11. We observe that removing the lite path leads to a significant drop in fidelity, as the model loses its primary information source. In this case, the model overly relies on the pretrained SD’s generative prior, often hallucinating textures, which results in perceptual artifacts and inflated non-reference scores like NIQE. On the other hand, removing the heavy path causes a moderate degradation in detail reconstruction, reflected by a noticeable drop in perceptual metrics such as LPIPS and DISTS.

A.9 Additional quantitative comparisons

Table 12 reports the results on the DRealSR dataset. CLIPQA leverages the pretrained image-text alignment model CLIP to evaluate the consistency between an image and a given textual prompt. This metric tends to favor results with richer texture details, regardless of their alignment with the ground truth.

While our method performs slightly worse than OSEDiff and AdcSR on this metric—both of which emphasize stronger generative capability—we achieve better results on perceptual fidelity metrics such as LPIPS and DISTS, which are more indicative of human-perceived visual quality. Importantly, all these improvements are achieved under significantly lower computational budgets, highlighting the efficiency and practicality of our approach.

A.10 Details of the Lite Decoder Structure

This section describes the architecture of the proposed Lite Decoder. In latent diffusion models, VAEs are typically employed to map images from pixel space to latent space, with reconstruction quality closely tied to the decoder’s design. To maintain model efficiency, we introduce structural simplification and channel pruning, aiming to balance lightweight design with high-fidelity reconstruction.

Table 12: Additional quantitative comparisons.

Metrics	FeMaSR	ResShift	OSDiff	AdcSR	PocketSR
CLIP-IQA \uparrow	0.5464	0.5342	0.6963	0.7049	0.6050
LPIPS \downarrow	0.3157	0.4006	0.2968	0.3046	0.2962
DISTS \downarrow	0.2239	0.2656	0.2165	0.2200	0.2139
FPS \uparrow	16.8	1.4	9.1	33.3	62.5

As illustrated in Figure 2, we simplify the original Stable Diffusion (SD) VAE decoder, which consists of five sequential modules—each typically composed of three ResBlocks, self-attention layers, and upsampling operations. Since the decoder performs $8\times$ spatial upsampling, all upsampling modules are retained. However, we remove the attention-equipped modules due to their high quadratic computational cost. To preserve reconstruction quality, a single ResBlock is placed after the final upsampling stage as the output head. Furthermore, we reduce the number of channels across the entire decoder, capping each layer at 64 channels.

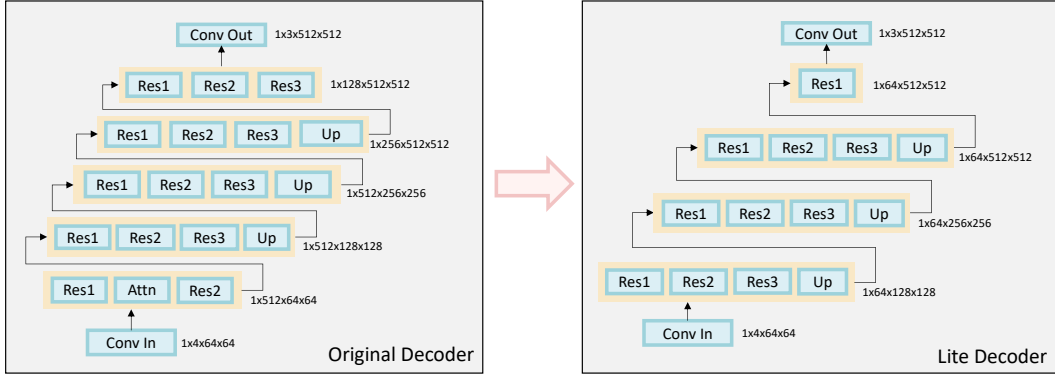


Figure 2: Structural comparison between the original SD VAE decoder and our Lite Decoder.

A.11 Details of Our Training Loss

Figure 3 illustrates the training pipeline of our model, along with the loss functions employed at each stage. The training process consists of two stages. In the first stage, we train the entire model end-to-end by replacing the original encoder-decoder with LiteED while keeping the U-Net intact. The loss used in this stage is a weighted sum of MSE, LPIPS, and adversarial losses:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{mse}} + \lambda_2 \mathcal{L}_{\text{lpips}} + \lambda_3 \mathcal{L}_{\text{gan}}, \quad (1)$$

where $\lambda_1 = 2$, $\lambda_2 = 2$, and $\lambda_3 = 0.25$.

In the second stage, LiteED is frozen, and we fine-tune the pruned U-Net. The fully trained U-Net from the first stage serves as the teacher, and multi-layer feature distillation is used to guide the pruned model. The original losses from the first stage are retained, with an additional distillation loss incorporated:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{mse}} + \lambda_2 \mathcal{L}_{\text{lpips}} + \lambda_3 \mathcal{L}_{\text{gan}} + \lambda_4 \mathcal{L}_{\text{distill}}, \quad (2)$$

where $\lambda_1 = 2$, $\lambda_2 = 2$, $\lambda_3 = 0.25$, and $\lambda_4 = 0.001$.

A.12 More Qualitative Comparisons

Figures 4 and 5 provide additional qualitative comparisons. As shown in Figure 4, PocketSR exhibits clear advantages in fidelity, particularly in recovering architectural details and regular structures. In contrast, other methods—especially multi-step approaches—often hallucinate content and diverge from the ground truth, despite their strong generative capabilities. Beyond its high fidelity, PocketSR also excels at fine texture generation, as illustrated in Figure 5, where it reconstructs high-quality textures for challenging elements such as starry skies, petals, wood, and rocks.

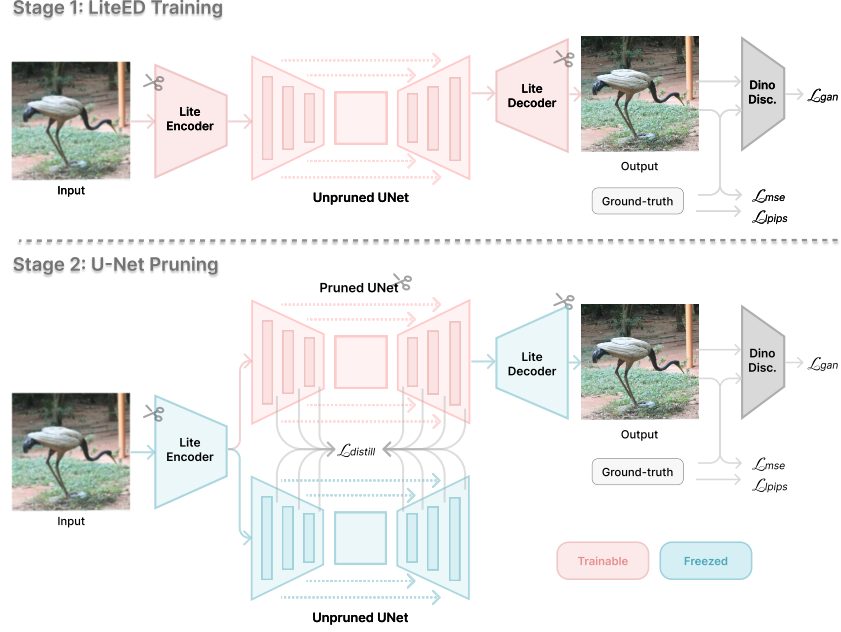


Figure 3: Loss composition at different training stages.

References

- [1] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, “Exploiting diffusion prior for real-world image super-resolution,” *IJCV*, 2024.
- [2] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, “Toward real-world single image super-resolution: A new benchmark and a new model,” in *ICCV*, 2019, pp. 3086–3095.
- [3] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-esrgan: Training real-world blind super-resolution with pure synthetic data,” in *ICCV*, 2021, pp. 1905–1914.
- [4] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, and C. Dong, “Diffbir: Toward blind image restoration with generative diffusion prior,” in *ECCV*. Springer, 2024, pp. 430–448.
- [5] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, “Seesr: Towards semantics-aware real-world image super-resolution,” in *CVPR*, 2024, pp. 25 456–25 467.
- [6] Z. Yue, J. Wang, and C. C. Loy, “Resshift: Efficient diffusion model for image super-resolution by residual shifting,” *NeurIPS*, vol. 36, pp. 13 294–13 307, 2023.
- [7] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen, “Sinsr: diffusion-based image super-resolution in a single step,” in *CVPR*, 2024.
- [8] R. Wu, L. Sun, Z. Ma, and L. Zhang, “One-step effective diffusion network for real-world image super-resolution,” *NeurIPS*, vol. 37, pp. 92 529–92 553, 2025.
- [9] B. Chen, G. Li, R. Wu, X. Zhang, J. Chen, J. Zhang, and L. Zhang, “Adversarial diffusion compression for real-world image super-resolution,” in *CVPR*, 2025.
- [10] J. Liang, H. Zeng, and L. Zhang, “Efficient and degradation-adaptive network for real-world image super-resolution,” in *ECCV*. Springer, 2022, pp. 574–591.
- [11] C. Chen, X. Shi, Y. Qin, X. Li, X. Han, T. Yang, and S. Guo, “Real-world blind super-resolution via feature matching with implicit high-resolution priors,” in *ACMMM*, 2022, pp. 1329–1338.

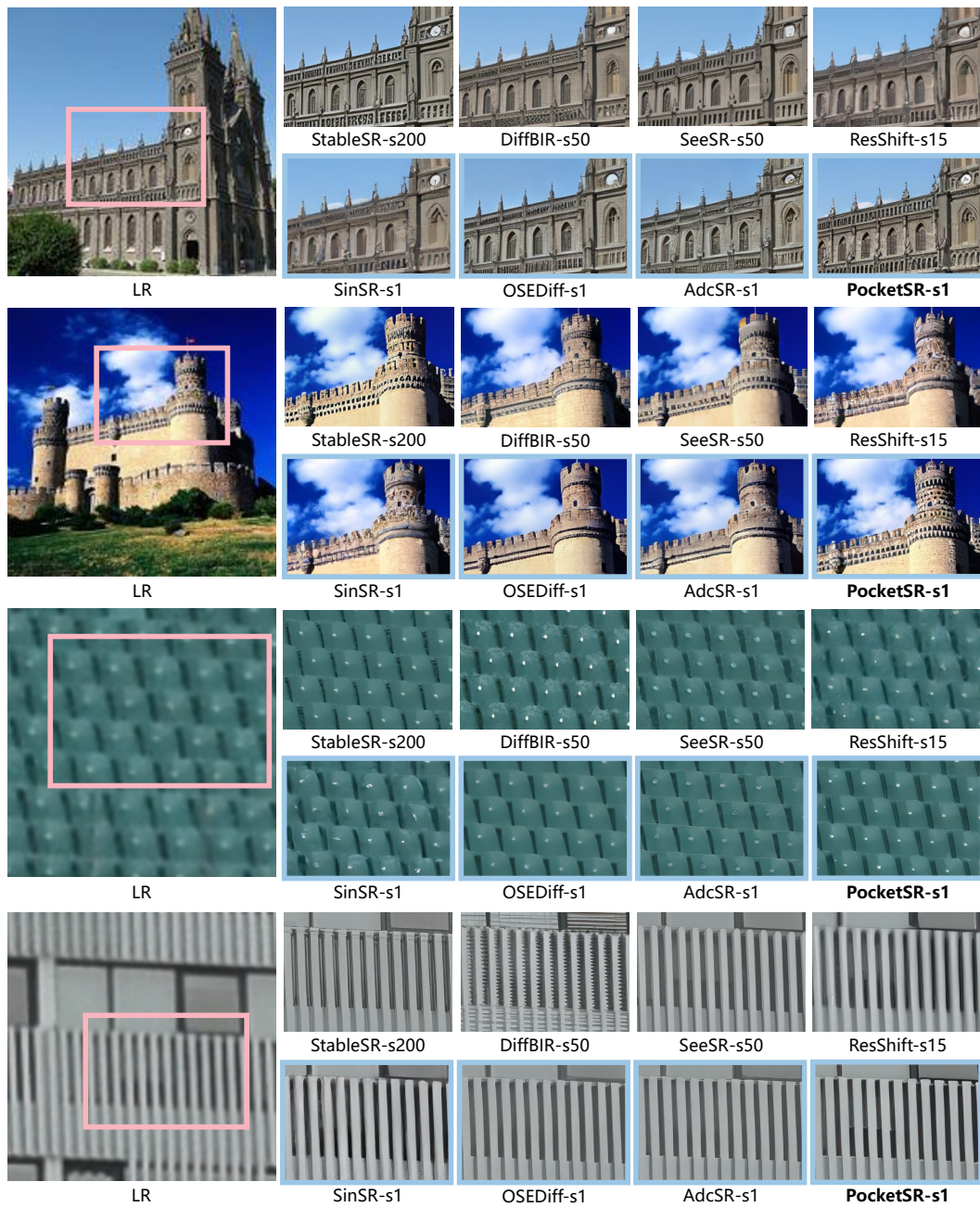


Figure 4: More qualitative results on real-world images. Single-step methods are highlighted for clarity. (1/2)

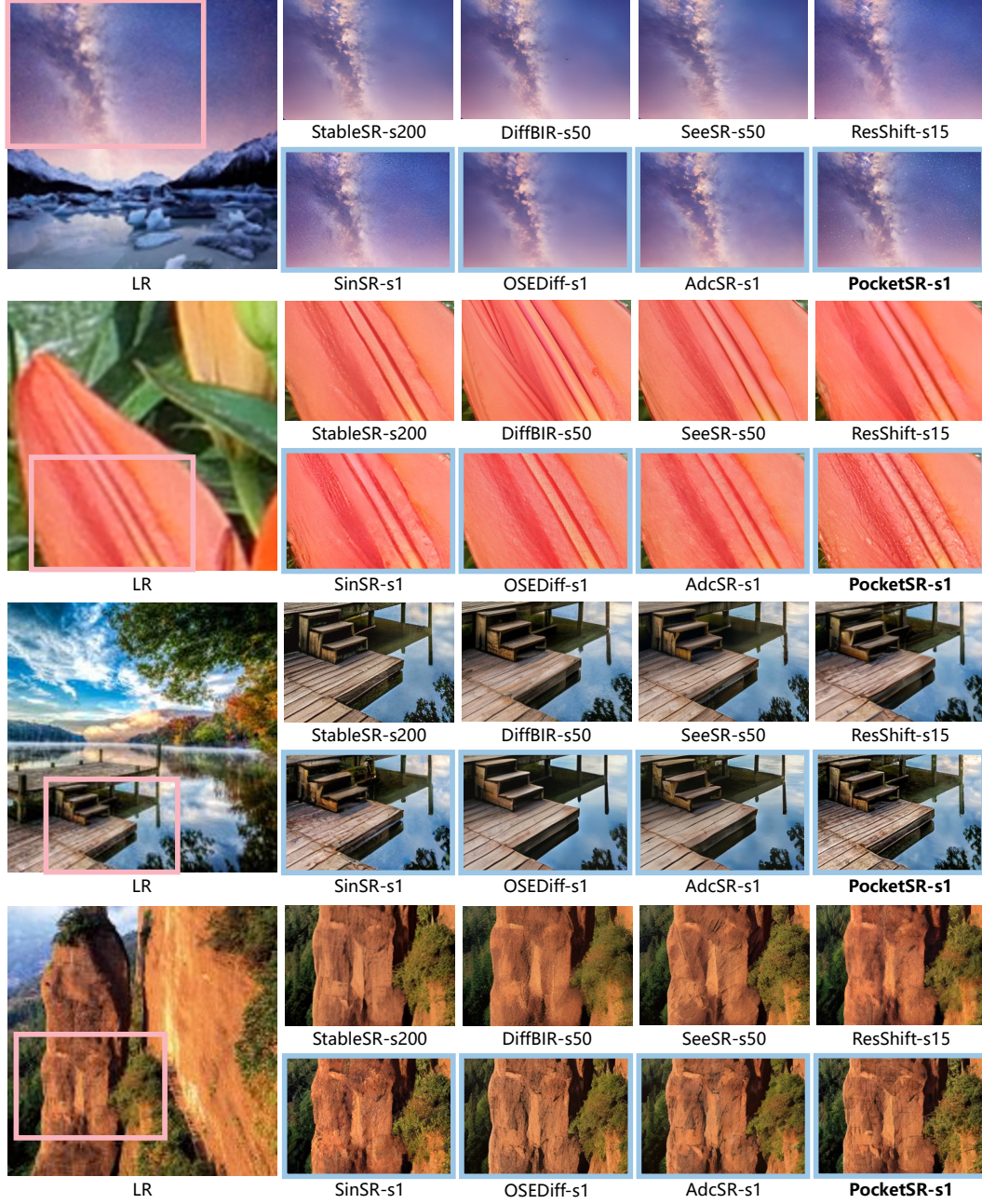


Figure 5: More qualitative results on real-world images. Single-step methods are highlighted for clarity. (2/2)