Acquire and then Adapt: Squeezing out Text-to-Image Model for Image Restoration

Supplementary Material

1. More Ablation Studies.

More Ablation of SE layers. To validate the necessity of multiple SE layers in FluxIR, we replaced them with a single MLP, meaning a single full-rank MLP was used to control all Flux MM-DiT blocks simultaneously. As shown in Tab. 1, the single MLP will strongly degrade the generation performance in all metric scores. Additionally, Fig. 1 illustrates that using a single MLP limits the model's generative capacity, resulting in lower-quality outputs. These findings highlight that the optimal design for our Flux adapter is to provide dedicated control for each Flux MM-DiT block.

Table 1. Comparison between Single MLP design and our multiple SE layers design on the *RealLQ250* dataset.

Control Layer	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑
FluxIR	**0.5639**	**70.78**	**0.6314**
Single MLP	0.5111	64.46	0.5547

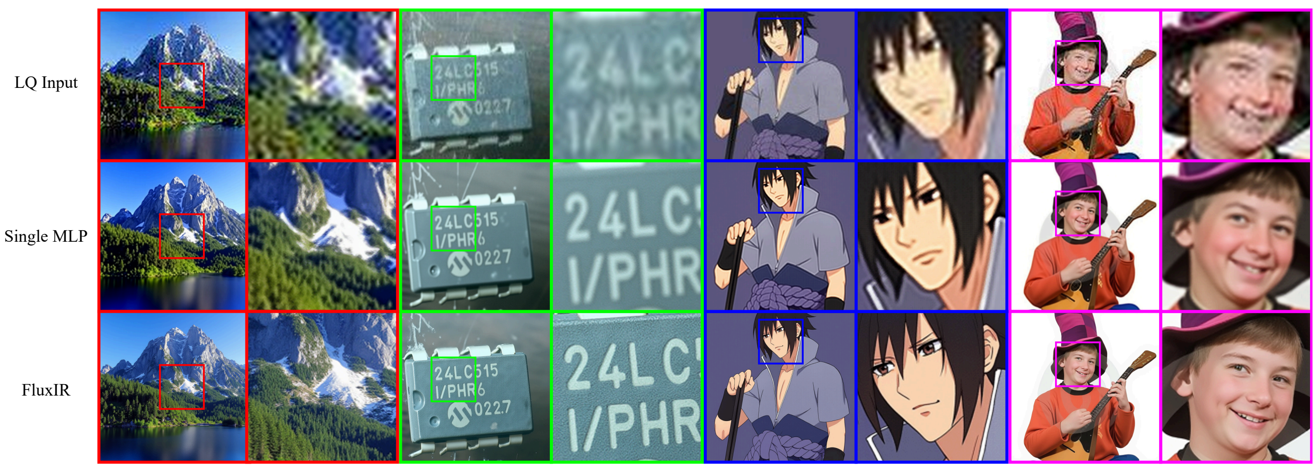

Figure 1. Visual results comparing the single MLP design and our multiple SE layers.

Ablation on multi-modality designs. In our proposed FluxIR, we introduce multi-modality controls on both the image and text information with a learnable T5 [9] embedding θ_p and a learnable CLIP [8] embedding θ_y. To justify the effectiveness of these designs, we evaluate the model variants by removing T5 embedding θ_y and CLIP embedding θ_y and text branch SE layer $\mathrm{SE}_p(\cdot)$, respectively. Tab. 2 presents the quantitative results on *RealLQ250* dataset. The results indicate that the text branch SE layer is crucial for enhancing the performance of our FluxIR model. The trainable T5 embedding θ_p and CLIP embedding θ_y show marginal differences from the baseline in evaluation metrics. As shown in Fig. 2, removing the text branch SE layers $\mathrm{SE}_p(\cdot)$ leads to a significant decline in image restoration performance. The trainable T5 embedding θ_p and CLIP embedding θ_y also contribute slight improvements in visual quality. The overall results demonstrate that the multi-modality design of FluxIR effectively boosts performance in the image restoration task.

Table 2. Ablation results of multi-modality designs on the *RealLQ250* dataset.

Multi-Modality	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑
Baseline	**0.5639**	**70.78**	**0.6314**
w/o θ_p,θ_y	0.5626	70.53	0.6308
w/o $\text{SE}_p(\cdot)$	0.5259	67.63	0.6266

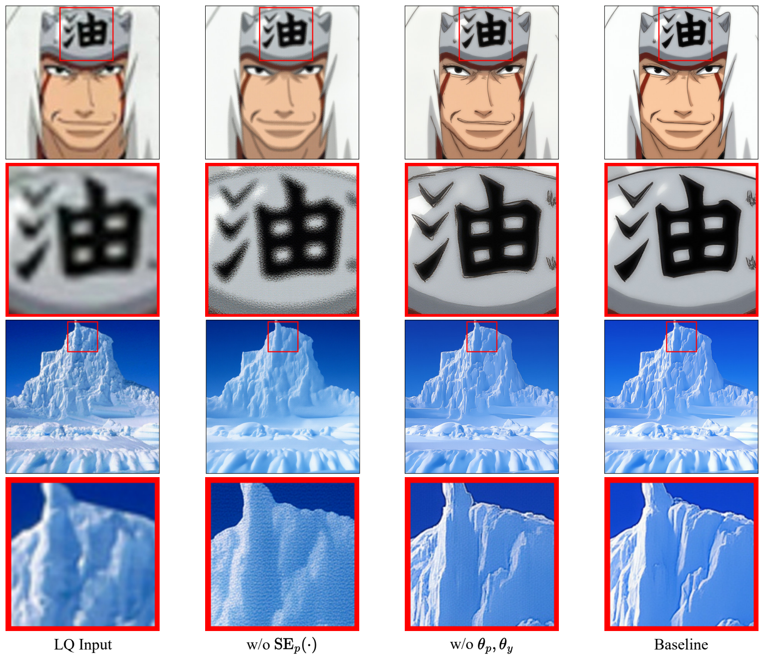

Figure 2. The visual comparisons of our multi-modality designs, *i.e*. text branch SE layers $\text{SE}_p(\cdot)$ and trainable embeddings θ_p, θ_y. Please zoom in for a better view.

2. Samples of Training dataset built by FluxGen

In this section, we present the dozens of samples produced by the FluxGen pipeline with the resolution of $1,024 \times 768$. Fig. 3 illustrates our generated training dataset obtained with an empty prompt, and demonstrates that an empty prompt is sufficient to produce diverse scene images with high resolution and aesthetic quality including cars, portraits, anime characters, animals, plants, food, buildings, indoor settings, furniture, the sea, and sunsets. We found that some ground truth images from FluxGen contain bokeh effects, which can occasionally cause localized blurriness in the restored results. However, based on subjective evaluations across four test datasets, the impact is minimal and acceptable. Similar issues could also arise in real-world datasets if not properly cleaned. Meanwhile, we show SDXL generated data in Fig. 4, which is also employed in the ablation studies. Without carefully designed prompt, SDXL cannot produce high-quality images for image restoration tasks. Fig. 5 shows more visual comparisons to further justify the effectiveness of FluxGen on the choice of text-to-image model and IQA selection. Furthermore, we generated 2,000 images from each of the five existing T2I models (PixelArt-Σ [4], Sana [11], SDXL [7], Playground [6], and Flux.1-dev [5]) for evaluation. As shown in Tab. 3, the images generated by our FluxGen pipeline achieved superior IQA scores.

Table 3. Comparisons with existing T2I generation methods.

Metric	PixArt-Σ	Sana	SDXL	Playground	Flux.1-dev	Ours
CLIPIQA ↑	0.4981	0.5135	0.5821	0.5822	0.6763	**0.7295**
MUSIQ ↑	64.93	66.75	70.49	70.35	75.02	**75.37**
MANIQA ↑	0.5519	0.5944	0.5831	0.6666	0.6590	**0.6962**

Figure 3. Samples of generated images by FluxGen pipeline.

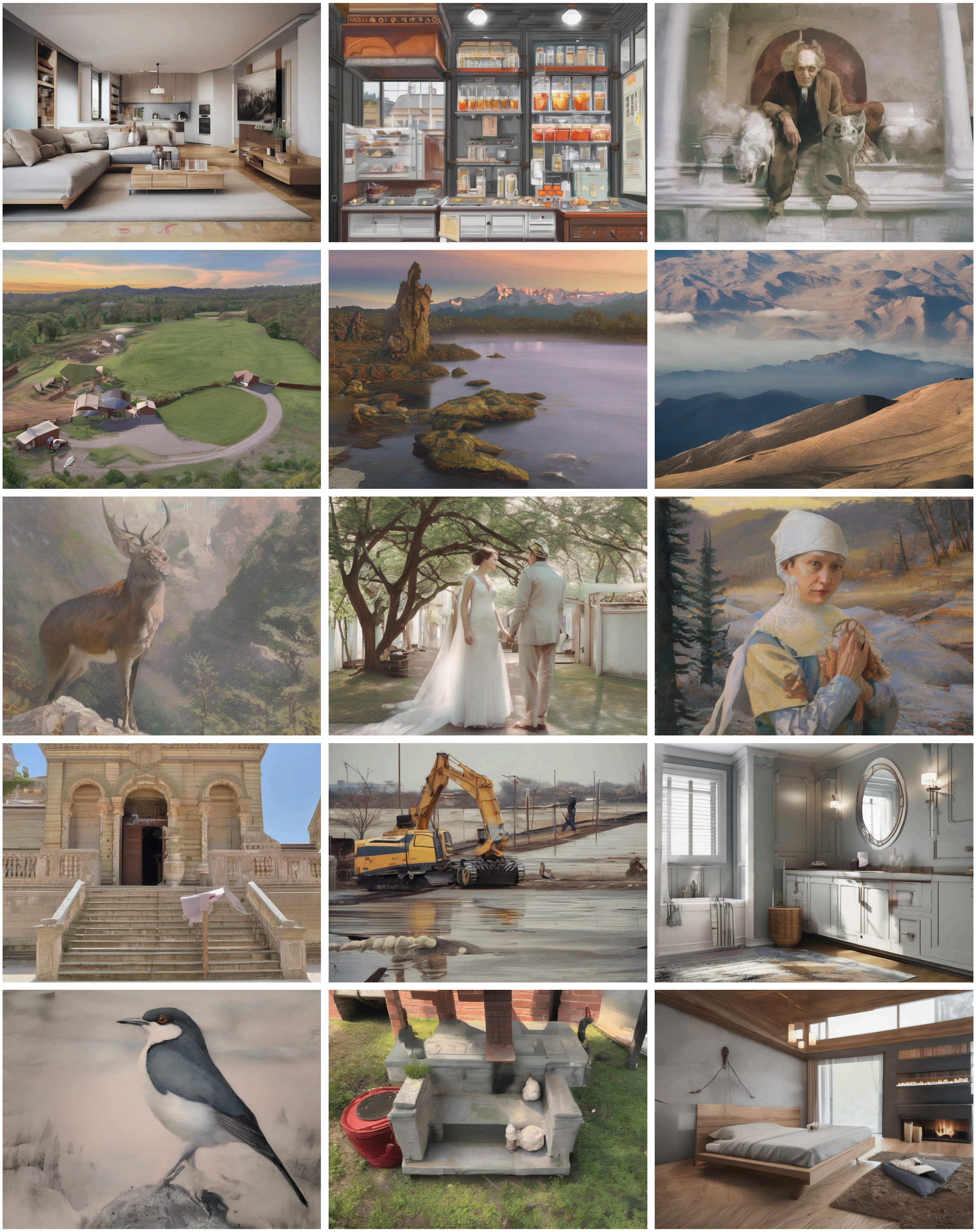

Figure 4. Samples generated by the SDXL.

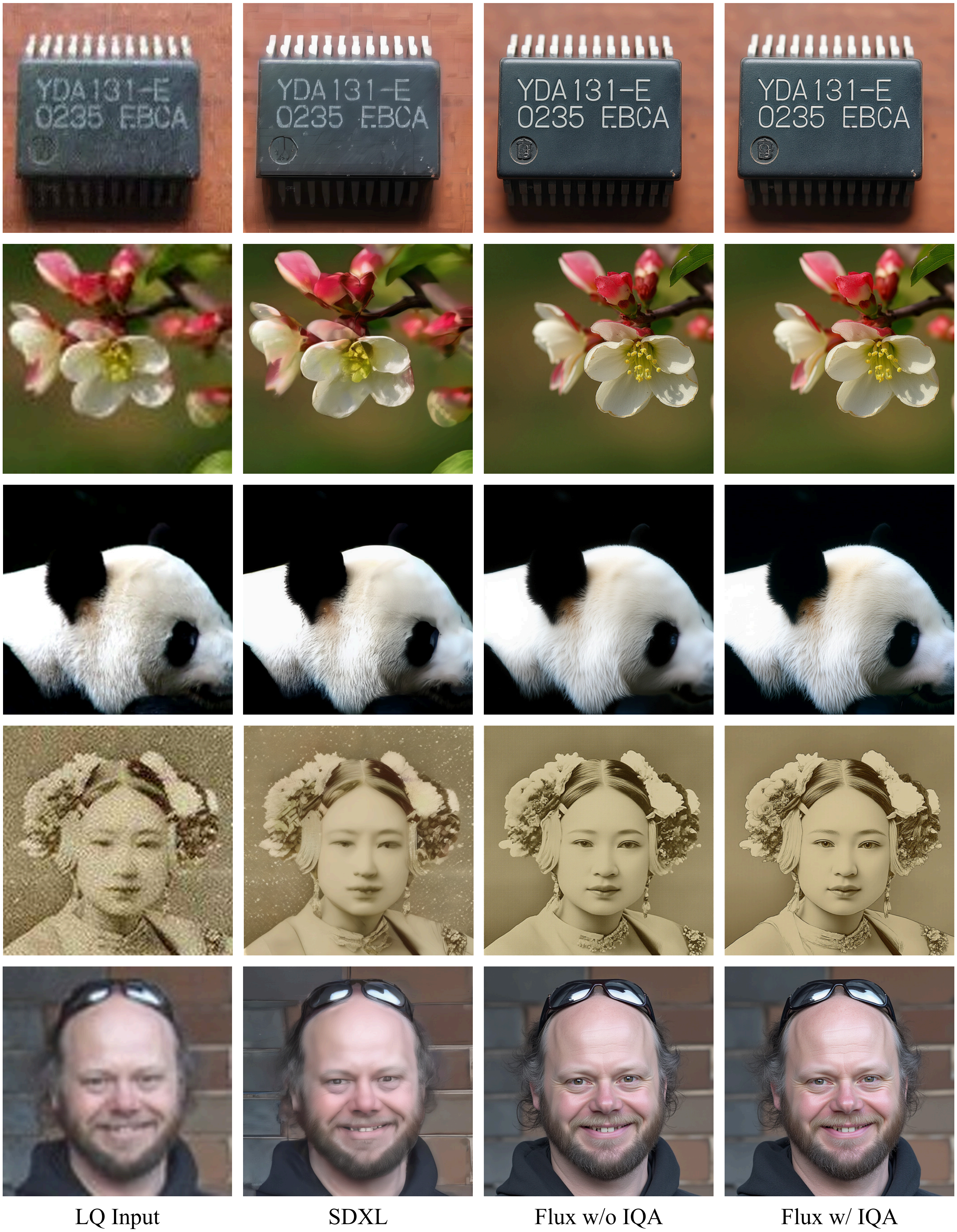

Figure 5. The visual comparisons of different FluxGen settings, where we study different T2I models, *i.e*. SDXL and Flux, and the usage of IQA selections. Please zoom in for a better view.

3. More Visualization Comparison.

Here, we provide additional visual results on synthetic and real-world datasets compared with state-of-the-art methods. Fig. 6 presents the visual results on the *DIV2K-Val* [1] dataset. Fig. 7 presents the visual results on the *RealSR* [3] dataset. Fig. 8 presents the visual results on the *DrealSR* [10] dataset. Fig. 9 presents the visual results on the *RealLQ250* [2] dataset. Our FluxIR achieves the best performance in terms of generation quality, texture details, and aesthetic quality. Please zoom in for a better view.

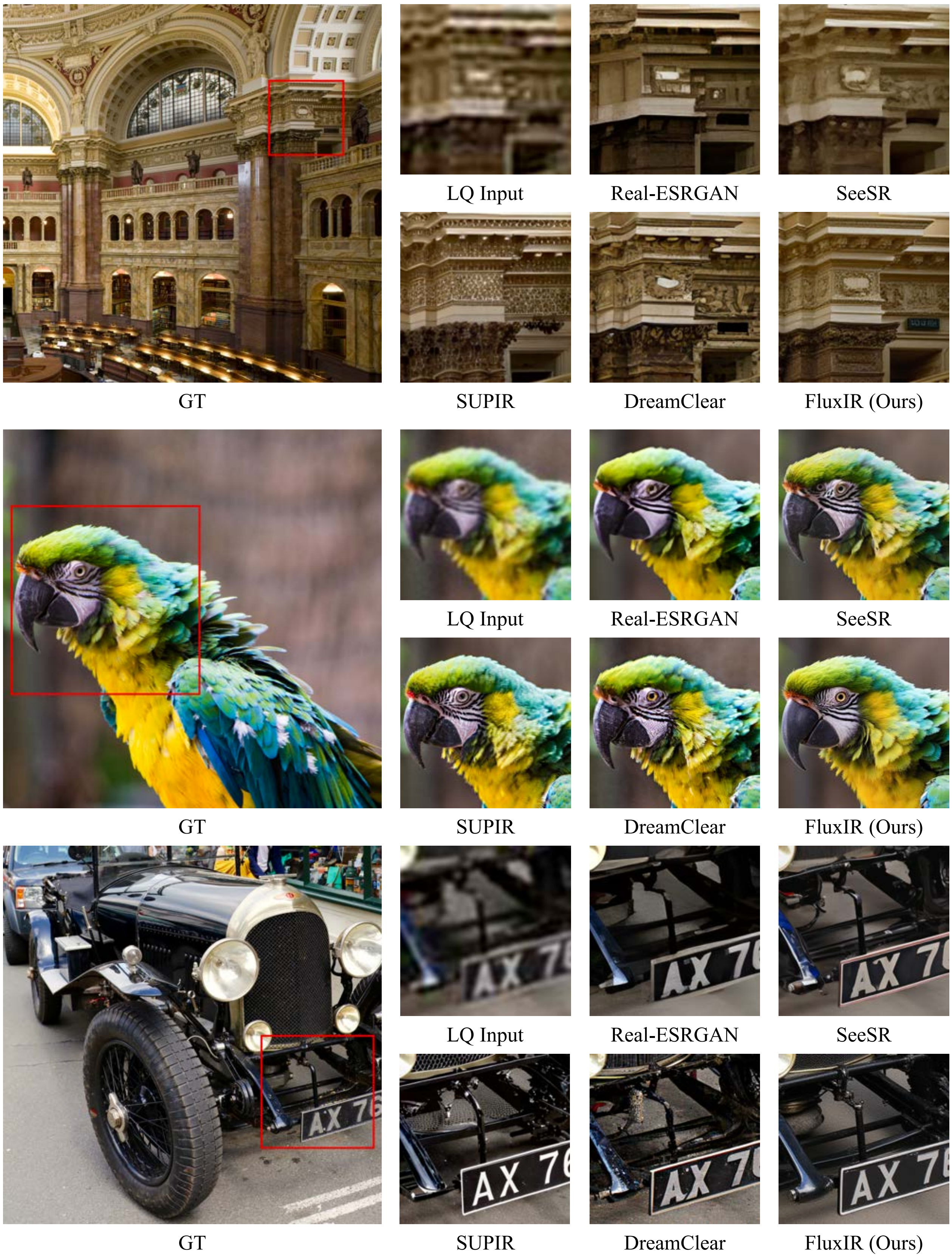

Figure 6. Visual comparison with SOTAs on *DIV2K-Val* dataset.

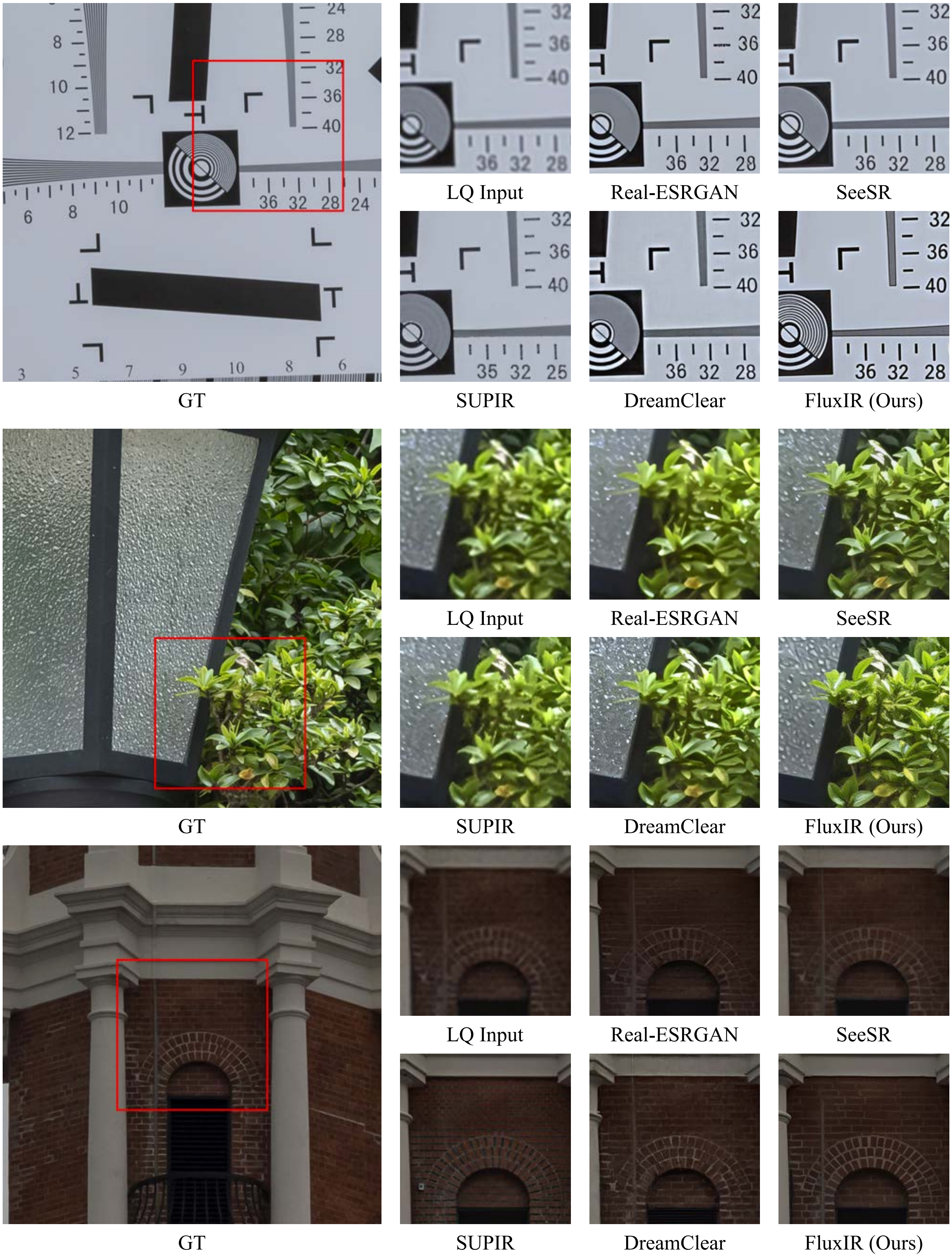

Figure 7. Visual comparison with SOTAs on *RealSR* dataset.

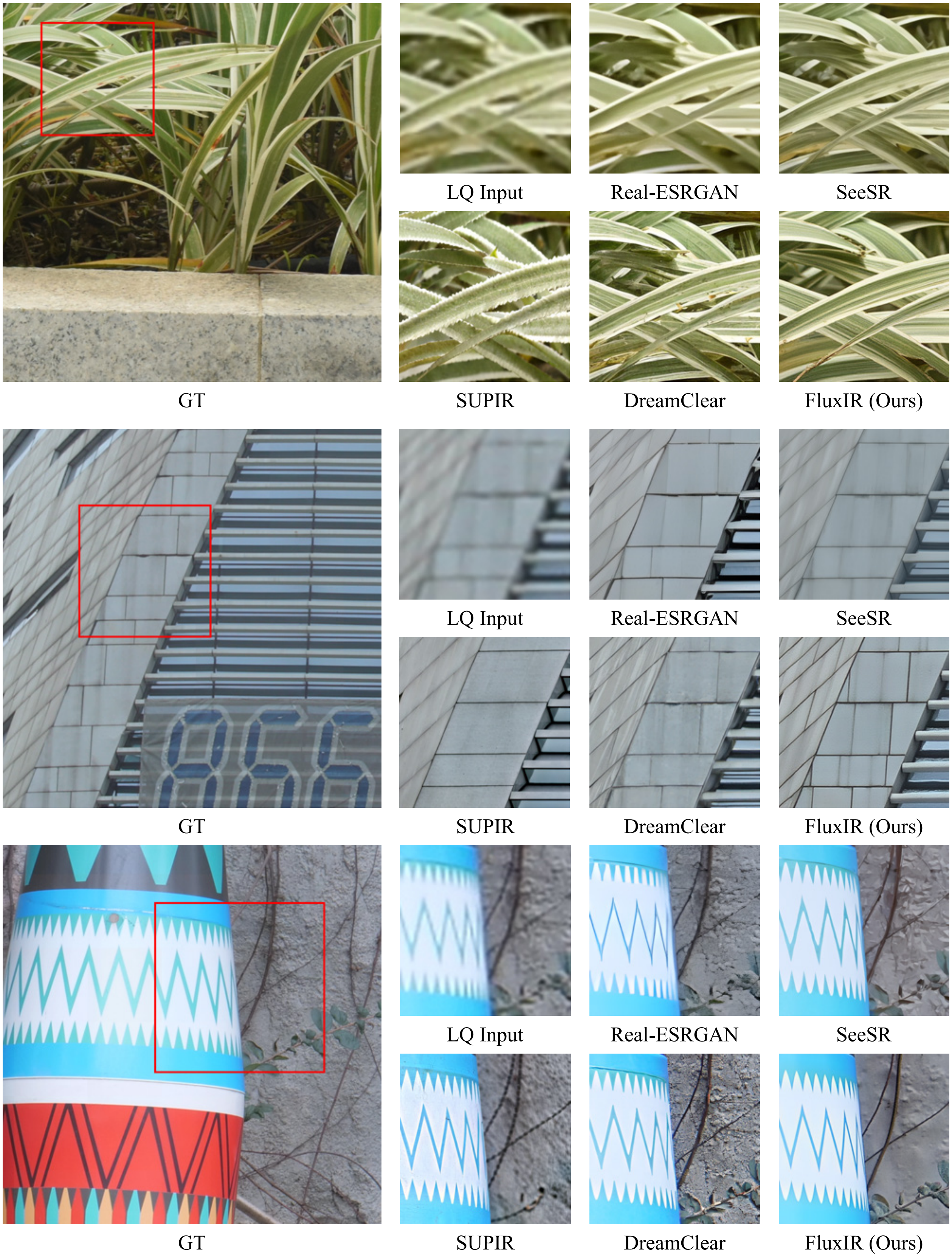

Figure 8. Visual comparison with SOTAs on *DrealSR* dataset.

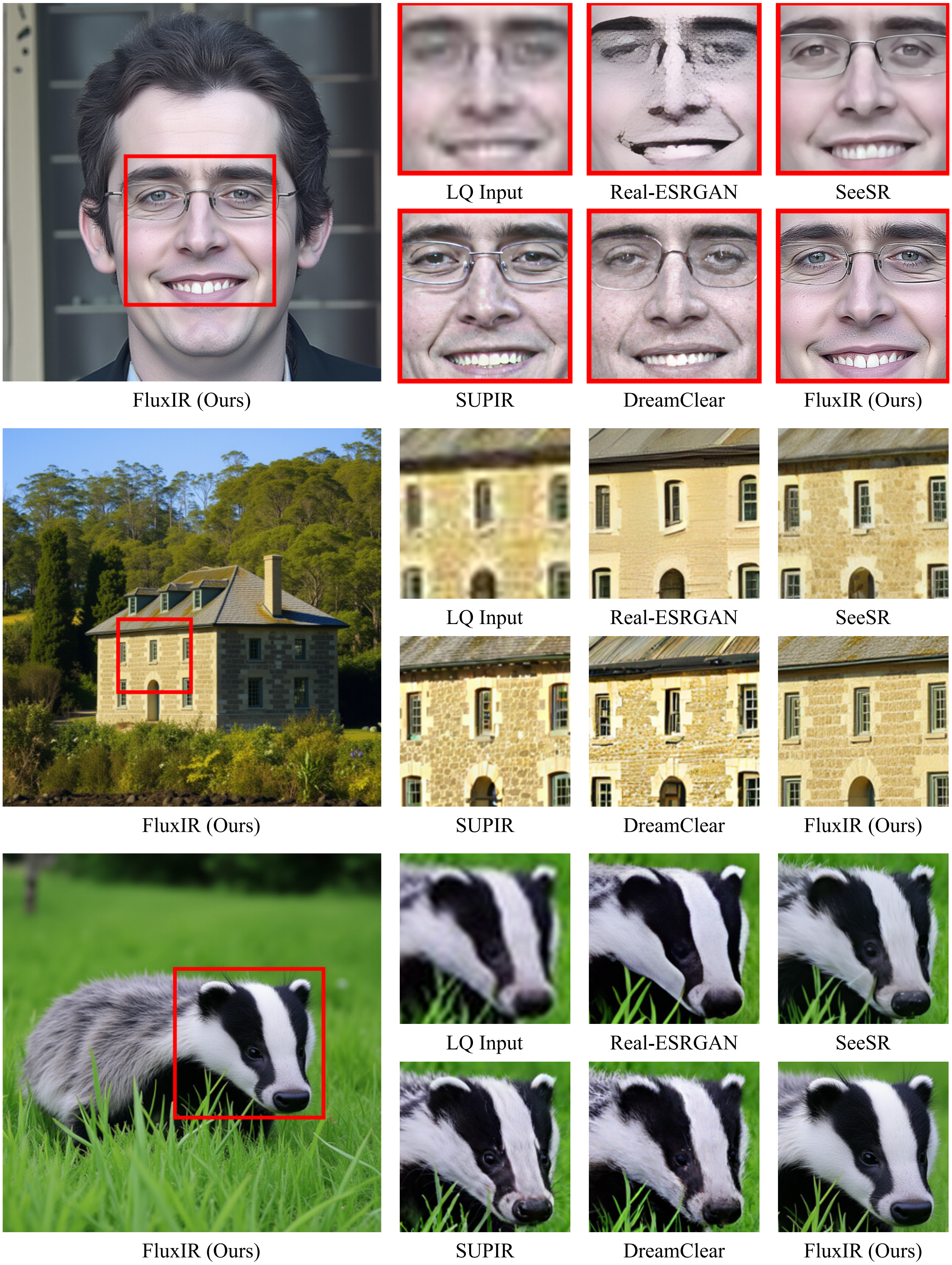

Figure 9. Visual comparison with SOTAs on *RealLQ250* dataset.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 6

[2] 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. *arXiv preprint arXiv:2410.18666*, 2024. 6

[3] 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*, pages 3086–3095, 2019. 6

[4] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, pages 74–91. Springer, 2024. 2

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

[6] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 2

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 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 *ECCV*, pages 101–117. Springer, 2020. 6

[11] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 2