DiT4SR: Taming Diffusion Transformer for Real-World Image Super-Resolution

——— *Supplemental Material* ———

Zheng-Peng Duan[1,2 *] Jiawei Zhang[2] Xin Jin[1] Ziheng Zhang[1] Zheng Xiong[2]
Dongqing Zou[2,3] Jimmy S. Ren[2,4] Chunle Guo[1,5] Chongyi Li[1,5 †]
[1]VCIP, CS, Nankai University [2]SenseTime Research [3]PBVR
[4]Hong Kong Metropolitan University [5]NKIARI, Shenzhen Futian
https://adam-duan.github.io/projects/dit4sr/

Our supplementary material gives more details about our method and more experimental results, which can be summarized as follows.

- We provide more related work in Section 1.
- We provide the implementation details in Section 2.
- We provide the detailed comparison with SD3-ControlNet in Section 3.
- We provide the details of ablation study in Section 4.
- We provide the details of user study in Section 5.
- We provide the PSNR and SSIM performance of our DiT4SR in Section 6.
- We provide the complexity comparison between our DiT4SR and other methods in Section 7.
- We provide the limitation of our DiT4SR and our future work in Section 8.
- We provide more qualitative comparisons in Section 9.

1. Related Work

Diffusion Transformer To enhance the generative capability of diffusion models, large-scale transformer architectures have been introduced, where diffusion transformer (DiT) [11] stands out. Building on DiT, large-scale T2I models, *e.g*. PixArt-α [5], SD3 [6], and Flux [2], are proposed. Specifically, SD3 and Flux leverage Multimodal Diffusion Transformers (MM-DiTs) to integrate text and image modalities through attention operation. In this way, the two modalities can fully interact, forming the core advantage of DiT. Our DiT4SR further enhances this advantage by incorporating the LR stream into the DiT blocks, enabling sufficient interaction between LR information and original features within the DiT blocks.

2. Implementation Details

Our DiT4SR is built upon Stable Diffusion 3.5, which shares a similar architecture with Stable Diffusion 3 [6]. We initialize the model parameters from SD3.5-Medium, and follow all the hyperparameter settings of SD3.5. Specifically, the total number N of the MM-DiT-Control blocks is 24. When the target resolution is set to 512×512, the height H, the width W, and the channel C of the noisy latent $\mathbf{Z} \in \mathbb{R}^{H \times W \times C}$ are 64, 64, and 16. Thus the length K and the dimension D of noisy image token $\mathbf{X} \in \mathbb{R}^{K \times D}$ and the LR image token $\mathbf{L} \in \mathbb{R}^{K \times D}$ are 1024 and 1536. The length M of the text token $\mathbf{C} \in \mathbb{R}^{M \times D}$ is set to 154. The training process is conducted on 512×512 resolution images with 8 NVIDIA 80G-A100GPUs. We train our model with a constant learning rate of $5e^{-5}$ with a batch size of 64. During inference, we adopt the default sampling schedule of SD3.5 with 40 sampling steps (T). The scale of classifier-free guidance (CFG) is set to 8 in our experiments. Following [1, 17], the prompt of the input LR image is obtained from LLaVA [9]. All the evaluation metrics are implemented by PyIQA [4]. Note that the metric of 'ClipIQA' is implemented with the setting of 'clipiqa+_vitL14_512' provided by PyIQA.

*This project is done during the internship at SenseTime Research.
†Corresponding author.

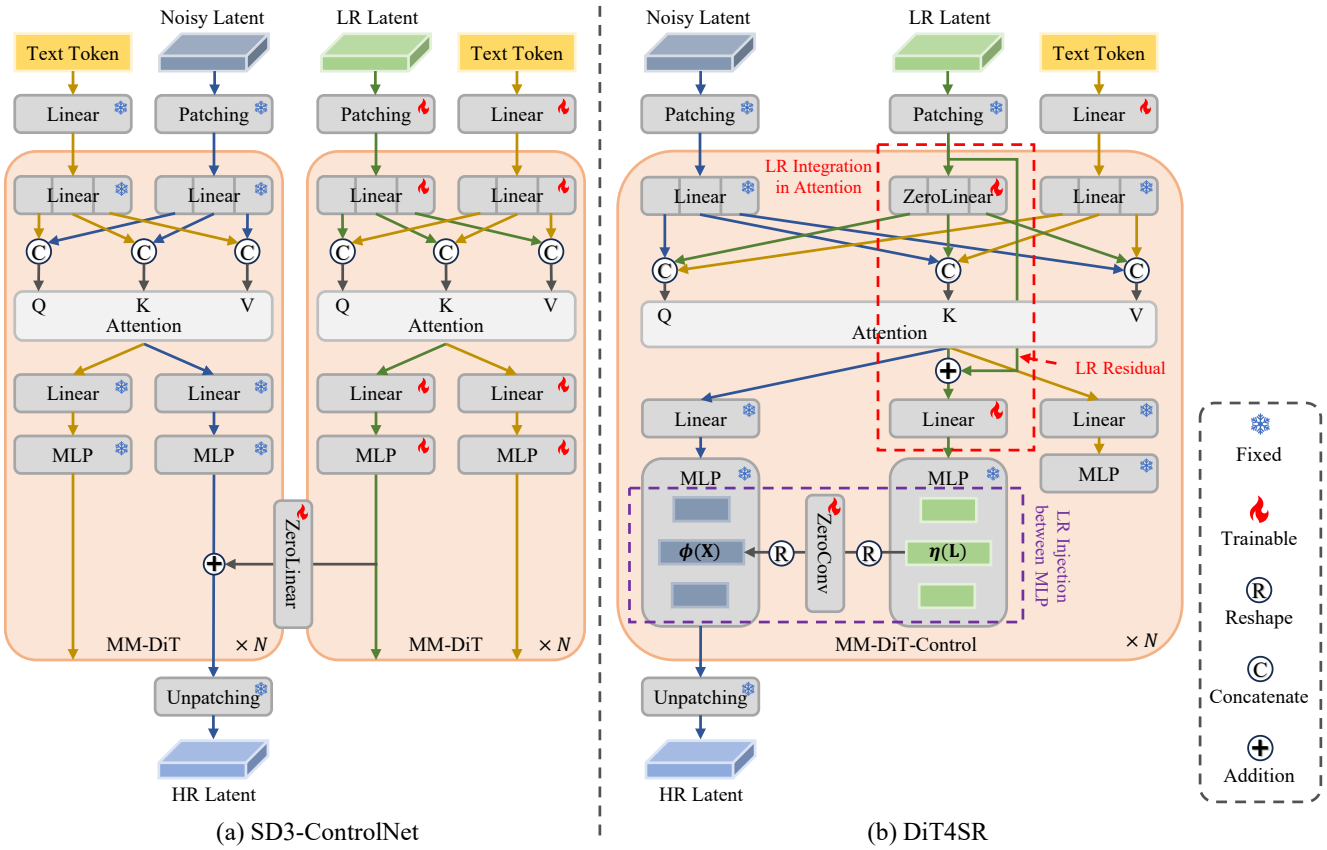


Figure 1. Detailed architecture comparison between SD3-ControlNet and our proposed DiT4SR. (a) SD3-ControlNet processes the LR Stream in additional MM-DiT blocks and injects LR information into the Noise Stream via trainable linear layers, establishing a one-way information flow. (b) Our DiT4SR directly integrates the LR Stream into the original DiT blocks, enabling bidirectional information flow through **LR Integration in Attention**. Additionally, **LR Injection Between MLP** incorporates convolutional layers to enhance local feature extraction, improving restoration fidelity.

3. Detailed Comparison with SD3-ControlNet

Figure 1 provides more detailed architecture comparison between SD3-ControlNet and our proposed DiT4SR. As illustrated in Figure 1, SD3-ControlNet processes the LR latent in additional MM-DiT blocks before being injected into the Noise Stream via trainable linear layers. This design establishes a one-way information flow from the LR Stream to the Noise Stream, limiting the information interaction between the two streams. The limited information interaction in SD3-ControlNet prevents the LR Stream from continuously adapting to the evolving state of the Noise Stream, hindering its ability to generate well-aligned guidance. This constraint potentially results in suboptimal guidance, affecting the quality of image restoration.

In contrast, our DiT4SR directly integrates the LR Stream into the original DiT blocks, allowing bidirectional information flow between the LR and Noise Streams. Specifically, we introduce **LR Integration in Attention**, where LR information is integrated directly into the attention computation, enabling continuous feature fusion in each block. Additionally, **LR Residual** is introduced to enhance the consistency of LR guidance throughout deeper transformer layers, mitigating the issue of diminishing influence. Furthermore, **LR Injection Beyond MLP** is incorporated through the convolutional layer to enhance local feature extraction, compensating for the weaker spatial capturing capability of DiT. By allowing the LR and Noisy Streams to evolve together, DiT4SR ensures that the LR guidance is progressively refined throughout the diffusion process, leading to more stable and high-fidelity restoration results compared to SD3-ControlNet.

We further visualize the LR Stream features at different depths for both SD3-ControlNet and our proposed DiT4SR. Following [10], the LR features of SD3-ControlNet is visualized with PCA in Figure 2 (b), and LR features of DiT4SR is in Figure 2 (d). In SD3-ControlNet, the LR features present unclear details and low edge distinction. This suggests that the one-way LR injection does not effectively preserve fine structures, resulting in weak guidance for the diffusion process. Consequently, as shown in Figure 2 (c), the restored image lacks sharpness, with indistinct edges and structural distortions.

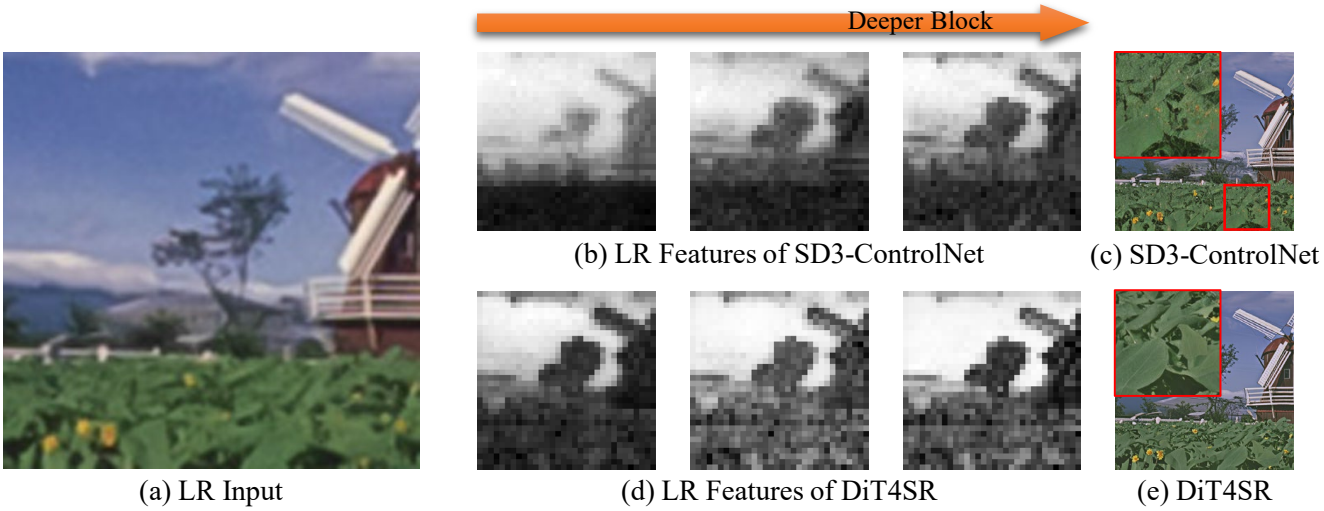


Figure 2. Comparison of LR feature evolution and final restoration results between SD3-ControlNet and our proposed DiT4SR. (a) The input low-resolution (LR) image. (b) and (d) are LR Stream features extracted from different depths of SD3-ControlNet and DiT4SR, which are visualized with PCA [10]. (c) and (e) are results of SD3-ControlNet and DiT4SR.

In contrast, our DiT4SR maintains clearer edge and less degradation across deeper layers (Figure 2 (d)), ensuring that fine-grained structures remain well-defined. This is enabled by our bidirectional information interaction, allowing the LR Stream to continuously refine itself based on the evolving Noise Stream. As a result, the final restoration (Figure 2 (e)) exhibits sharper edges and more distinct textures.

4. Details of Ablation Study

In Figure 3, we provide the detailed architectures of the four variants and the full model in the ablation study.

Variant A. Variant **A** removes the LR Stream from the attention operation. Instead, the LR Stream calculates the attention via self-attention mechanism, which can be formulated as

$$\text{Attention}(P_{\mathbf{Q}}^{\mathbf{L}}(\mathbf{L}), P_{\mathbf{K}}^{\mathbf{L}}(\mathbf{L}), P_{\mathbf{V}}^{\mathbf{L}}(\mathbf{L})) = \underbrace{\text{softmax}(\frac{P_{\mathbf{Q}}^{\mathbf{L}}(\mathbf{L}) P_{\mathbf{K}}^{\mathbf{L}}(\mathbf{L})^T}{\sqrt{d}})}_{\text{attention map}} P_{\mathbf{V}}^{\mathbf{L}}(\mathbf{L}), \tag{1}$$

where $P_{\mathbf{Q}}^{\mathbf{L}}$, $P_{\mathbf{K}}^{\mathbf{L}}$, and $P_{\mathbf{V}}^{\mathbf{L}}$ are trainable linear projections for LR image token $\mathbf{L}$. The LR Residual and LR Injection between MLP are preserved.

Variant B. Variant **B** removes the LR Residual, preventing the LR Stream from maintaining a direct pathway across deeper layers. LR Integration in Attention and LR Injection between MLP are preserved.

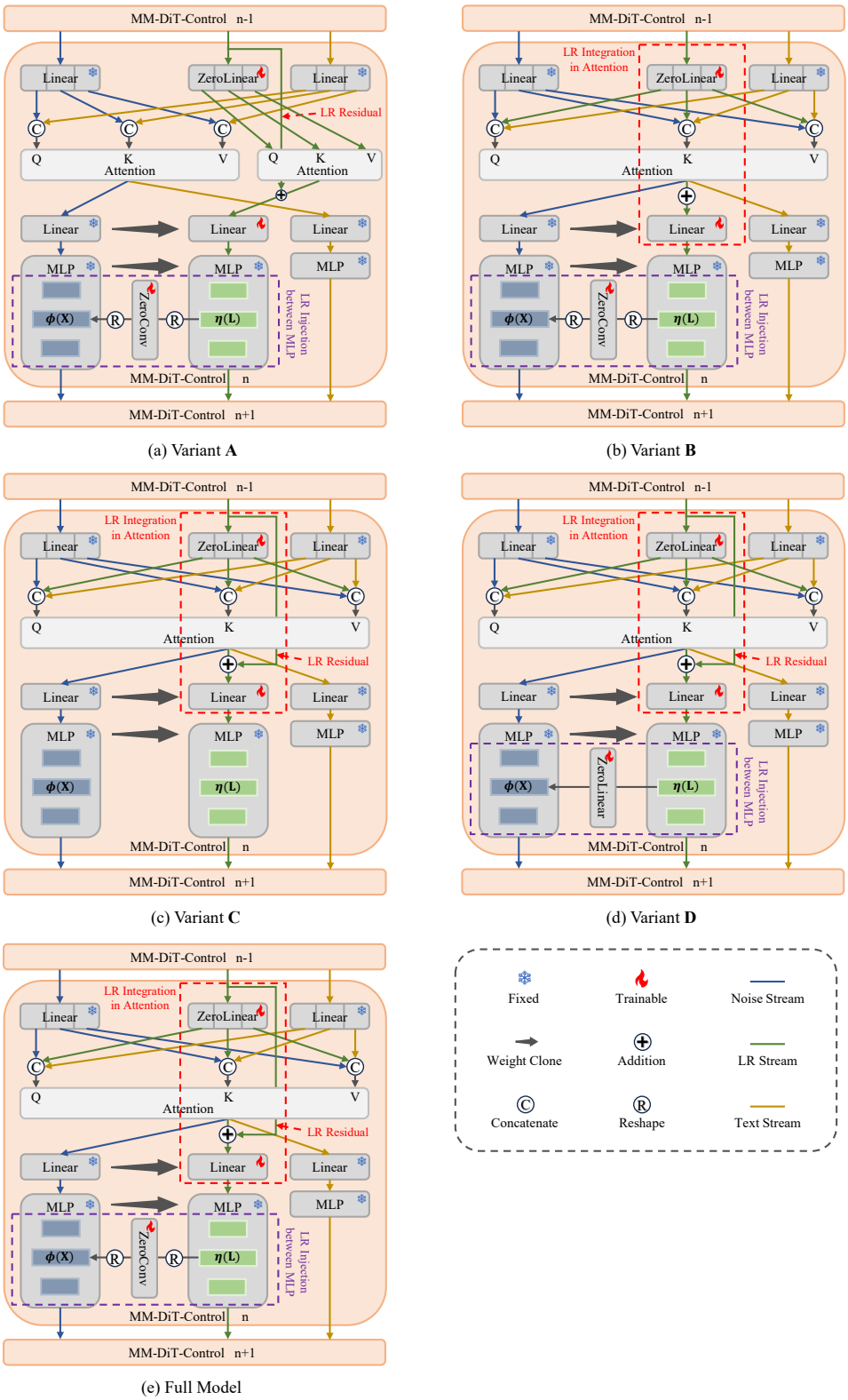
Variant C. Variant **C** removes the LR Injection between MLP, meaning LR information is only integrated into the model via attention mechanism. LR Integration in Attention and LR Residual are preserved.

Variant D. Variant **D** replaces the 3×3 depth-wise convolution in LR Injection between MLP with a linear layer. LR Integration in Attention and LR Residual are preserved.

Full Model. The full model retains all three components: LR Integration in Attention, LR Residual, and LR Injection between MLP, ensuring bidirectional information exchange and a comprehensive interaction between the LR and Noise Streams. This configuration allows the LR Stream to evolve alongside the Noise Stream, facilitating adaptive and consistent guidance throughout the diffusion process. Meanwhile, the 3×3 depth-wise convolution layer in LR Injection also compensates for the limited local information-capturing ability of DiT.

5. Details of User Study

Figure 4 presents an example of our user study. In the user study, volunteers are provided with two restoration results, where one is from our method, and the other is from the compared methods. The volunteers are asked the following questions: 1) Which restoration result has higher image realism? 2) Which restoration result has better fidelity to the original image content?



(a) Variant **A**

(b) Variant **B**

(c) Variant **C**

(d) Variant **D**

(e) Full Model

Figure 3. Variant **A** (a), **B** (b), and **C** (c) remove the LR Integration in Attention, LR Residual, and LR Injection between MLP, respectively. Variant **D** (d) replaces the convolution layer with the linear layer in LR Injection. (e) is the full model of our DiT4SR.

User Study

Here is a low-resolution (LR) input image. We aim to recover its high-resolution (HR) version.

The following two images are the restoration results of different methods.

* Which restoration result has higher image realism?

- ○ Left image
- ○ Right image

* Which restoration result has better fidelity to the original image content?

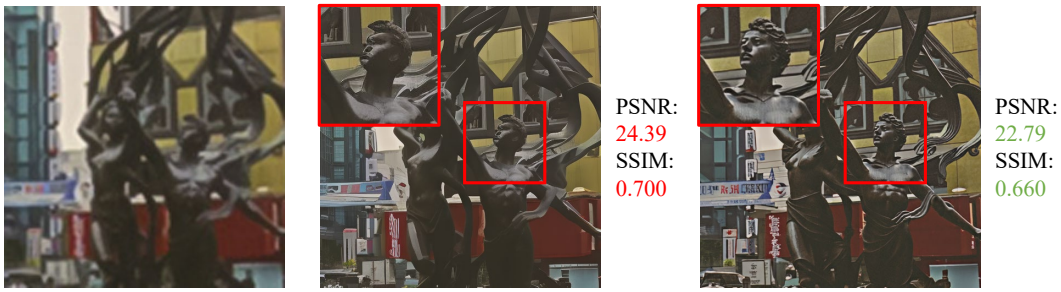
- ○ Left image
- ○ Right image

Figure 4. One comparison example in the user study. In each comparison, the volunteers are asked the following questions: 1) Which restoration result has higher image realism? 2) Which restoration result has better fidelity to the original image content?

6. PSNR and SSIM Performance

Datasets	Metrics	Real-ESRGAN	SwinIR	ResShift	StableSR	SeeSR	DiffBIR	OSEDiff	SUPIR	DreamClear	SD3-ControlNet	DiT4SR
DrealSR	PSNR ↑	28.615	28.497	28.692	28.407	28.074	25.929	27.915	24.988	28.394	27.205	25.806
	SSIM ↑	0.805	0.804	0.787	0.795	0.768	0.652	0.783	0.647	0.745	0.734	0.682
RealSR	PSNR ↑	25.686	26.308	26.387	23.437	25.188	24.240	25.148	23.679	24.847	24.004	23.378
	SSIM ↑	0.761	0.773	0.756	0.692	0.722	0.665	0.734	0.664	0.699	0.685	0.664

Table 1. PSNR and SSIM performance of state-of-the-art Real-ISR methods on two real-world benchmarks. Best and second best performance are highlighted in red and blue, respectively.



(a) Bicubic Input (b) SeeSR (c) DiT4SR

Figure 5. (a) LR input. (b) SeeSR produces higher PSNR (24.39) and SSIM (0.700) but exhibits over-smoothed textures and lacks fine-grained details. (c) Our DiT4SR achieves lower PSNR (22.79) and SSIM (0.660) but reconstructs richer details and sharper textures, demonstrating improved perceptual quality despite lower full-reference metrics.

In Table 1, we report the PSNR and SSIM performance of state-of-the-art Real-ISR methods on two real-world benchmarks, including our DiT4SR. Although our DiT4SR can generate richer details and achieve better visual effects, DiT4SR shows no advantage on these full-reference metrics (PSNR and SSIM), which is further demonstrated by Figure 5. This can be attributed to the limitations of these full-reference metrics, which is also mentioned in previous studies [3, 7, 17]. Therefore, we argue that comparing PSNR and SSIM performance across different methods is not particularly meaningful for the Real-ISR task.

7. Parameters and Inference Time

Methods	Base Model	Params	Sample Steps	Inference Time
ResShift	Diffusion	16.7M	15	0.79s
StableSR	SD2	1409.1M	200	13.18s
SeeSR	SD2	2283.7M	50	5.14s
DiffBIR	SD2	1716.7M	50	4.12s
SUPIR	SDXL	4801.2M	50	11.85s
SD3-ControlNet	SD3.5-Medium	3504.4M	40	4.35s
DiT4SR	SD3.5-Medium	2716.8M	40	5.61s

Table 2. Complexity comparison between different methods. All evaluations are conducted on an NVIDIA 80G-A100 GPU, where each method generates 512×512 results from 128×128 inputs.

Table 2 presents a comparison of different methods in terms of model parameters, sampling steps, and inference time. All evaluations are conducted on an NVIDIA 80G-A100 GPU, where each method generates 512×512 results from 128×128 inputs.

ResShift [18], utilizing a lightweight diffusion-based model, has the lowest computational cost, requiring only 0.79s per inference with 15 sampling steps. StableSR [14] is significantly more computationally expensive, requiring 200 sampling steps and 13.18s for inference. SeeSR [15] and DiffBIR [8] employ 50 sampling steps, with DiffBIR achieving a slightly faster inference time. SUPIR [17], leveraging SDXL [12], has the largest model size (4801.2M params) and requires 11.85s per inference, reflecting the increased computational demand of scaling to SDXL. Both SD3-ControlNet [13, 19] and our DiT4SR are built upon SD3.5-Medium [6]. They both adopt 40 sampling steps, with DiT4SR having fewer parameters but a slightly higher inference time.

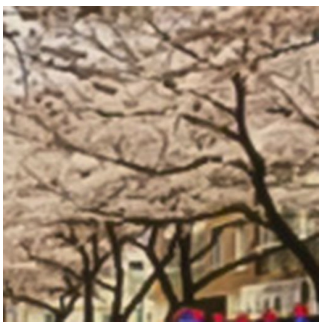
8. Limitation and Future Work

Like most diffusion-based Real-ISR methods [1, 15–17], our DiT4SR also requires a detailed prompt describing the image content as input. During inference, some approaches, such as SUPIR [17] and DreamClear [1], first pass the LR image through a degradation removal model before using LLAMA to generate the prompt. However, we observe that degradation removal may erase or alter the original information in the LR image, leading to hallucinated content in the generated prompt, such as "snow-covered trees" and "black and white style" in Figure 6 (b).

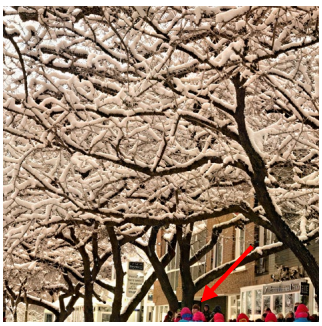
In our work, we choose to use the LR image as input to LLAMA directly, preserving more of the original content. Although this approach can somewhat reduce hallucinations in certain cases, it still cannot guarantee the generation of entirely accurate prompts, which is shown in Figure 6 (d). In Figure 6 (f) and (h), both SUPIR and DiT4SR hallucinate an airplane in the background, resulting in restoration errors that deviate from the ground truth. Therefore, this highlights the challenge of accurately extracting prompts from LR images, which remains an open research problem. We leave this as our future research direction.

9. More Qualitative Comparisons

Figure 7, Figure 8, Figure 9, and Figure 10 provide more qualitative comparisons on four datasets (DrealSR, RealSR, RealLQ250, RealLR200). Our method is capable of generating results with richer details than the compared methods while simultaneously maintaining high fidelity.

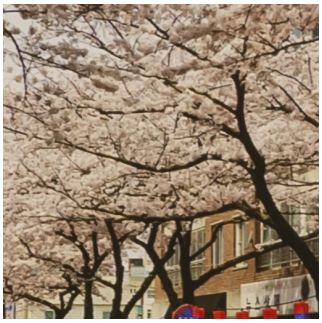


(a) Bicubic Input

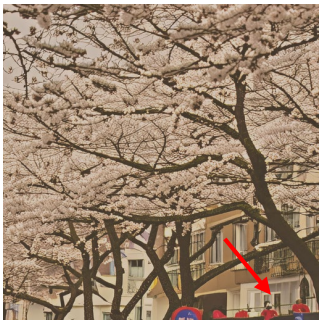


(b) SUPIR

SUPIR Prompt:
The image features a snow-covered tree with a group of people walking underneath it. The tree is surrounded by several other trees, creating a picturesque winter scene. The people are walking in a line, with some closer to the tree and others further away. The image is a black and white photograph, which adds a timeless and classic feel to the scene. The snow-covered trees and the people walking underneath them create a sense of tranquility and beauty, capturing the essence of a winter day.



(c) Ground Truth



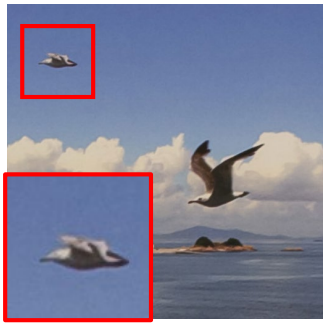
(d) DiT4SR

DiT4SR Prompt:
The image features a tree-lined street with a row of trees on both sides. The trees are covered in white flowers, creating a beautiful and serene atmosphere. A group of people are walking underneath the trees. The street is lined with buildings, and there are several windows visible on the buildings. The overall scene is picturesque and inviting.

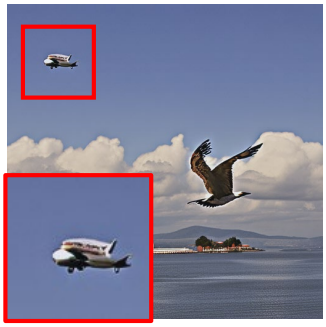
(e) Bicubic Input

(f) SUPIR

SUPIR Prompt:
The image features a bird flying over a body of water, possibly an ocean, with a mountainous background. The bird is captured in mid-flight, soaring gracefully through the sky. The scene is depicted in a black and white style, giving it a timeless and artistic quality. In addition to the bird, there is a small airplane flying in the sky, adding another element of interest to the scene. The combination of the bird, the ocean, and the mountainous landscape creates a serene and picturesque atmosphere.



(g) Ground Truth



(h) DiT4SR

DiT4SR Prompt:
The image features a large bird flying over a body of water, possibly a lake or a bay. The bird is soaring high in the sky, with its wings spread wide. The water below is calm, and the scene is serene. In the background, there is a small airplane flying at a higher altitude, adding an interesting contrast to the bird's flight. The airplane is positioned towards the left side of the image, while the bird is flying towards the right side.

Figure 6. (a) and (e) show the bicubic input LR images. (c) and (g) show the ground truth images. (b) and (f) present the outputs from SUPIR, where prompts generated using LLAMA after degradation removal introduce hallucinated contents, such as “snow-covered trees”, “black and white style” and “airplane in the background”. These inaccuracies lead to incorrect image restoration. (d) and (h) display the outputs from DiT4SR, where prompts are generated directly from the LR input without degradation removal, somewhat reducing hallucinations but still suffering from incorrect contents (e.g., the hallucinated people and airplane). The hallucinated contents in prompts are marked in red.

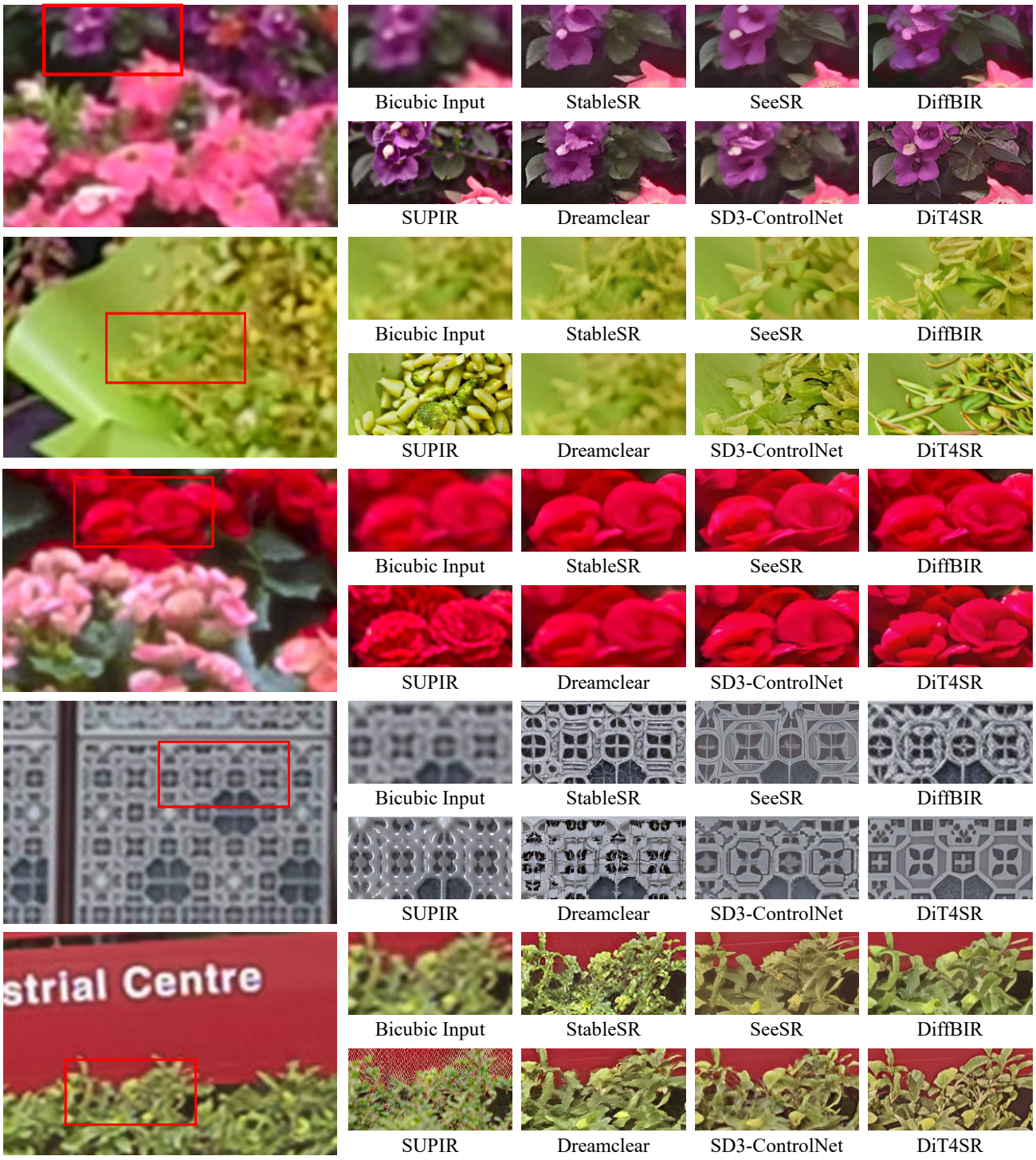


Figure 7. More qualitative comparisons with state-of-the-art Real-ISR methods on DrealSR (the first two rows) and RealSR (the last three rows). Our DiT4SR achieves the best performance in terms of image realism and detail generation while maintaining fidelity to the input LR image.

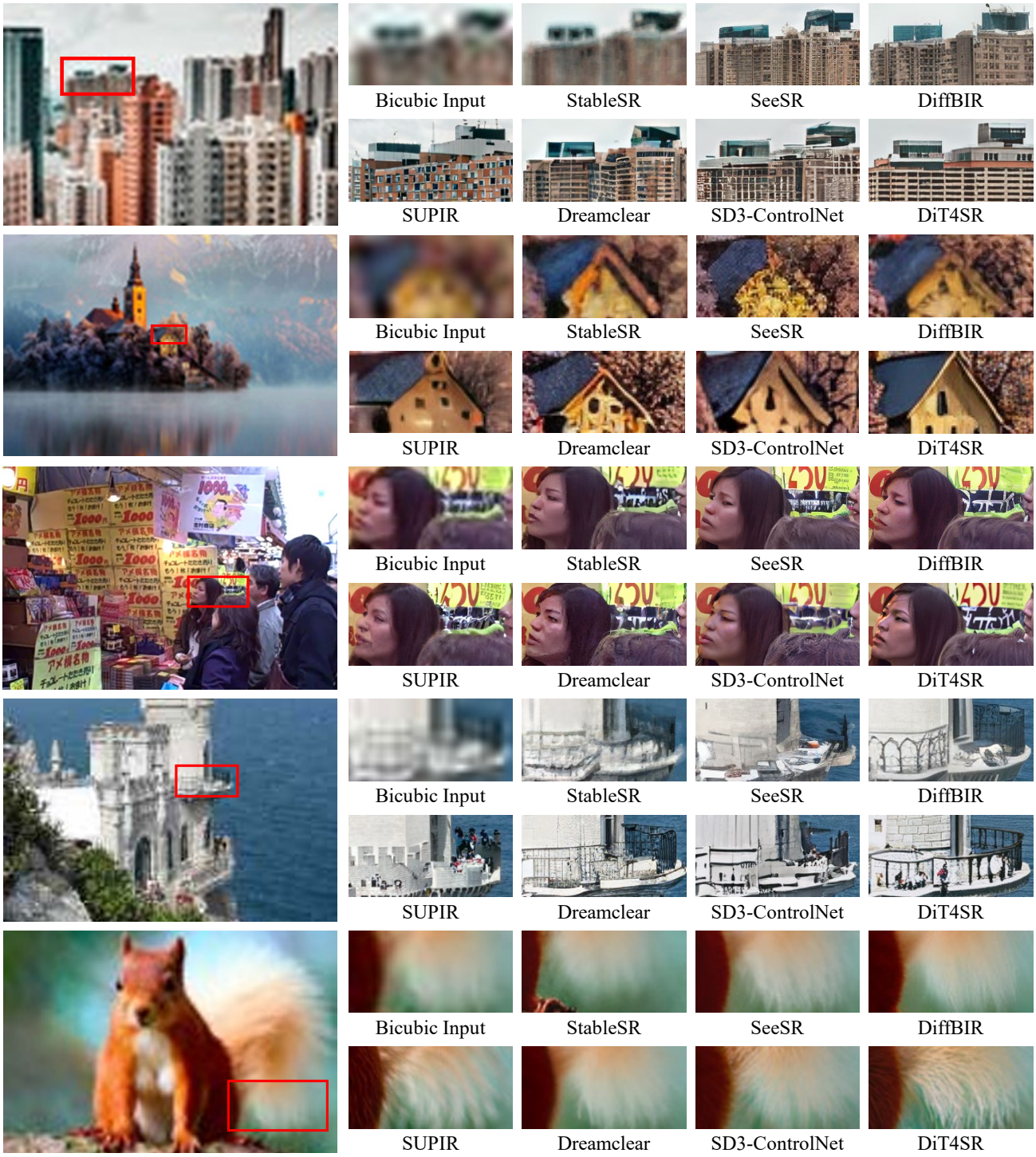


Figure 8. More qualitative comparisons with state-of-the-art Real-ISR methods on RealLQ250 (the first two rows) and RealLR200 (the last three rows). Our DiT4SR achieves the best performance in terms of image realism and detail generation while maintaining fidelity to the input LR image.

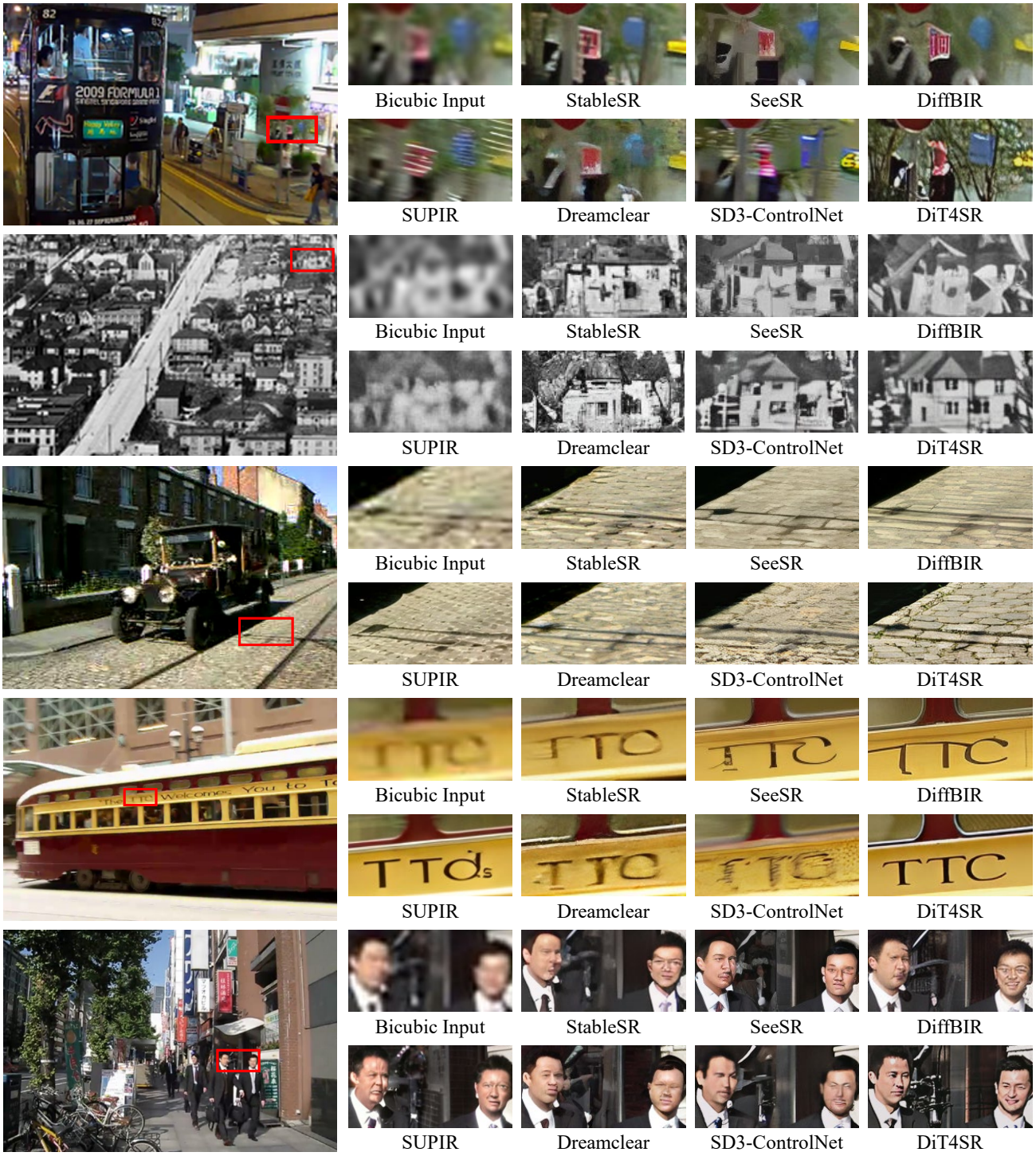


Figure 9. More qualitative comparisons with state-of-the-art Real-ISR methods on RealLQ250 (the first two rows) and RealLR200 (the last three rows). Our DiT4SR achieves the best performance in terms of image realism and detail generation while maintaining fidelity to the input LR image.

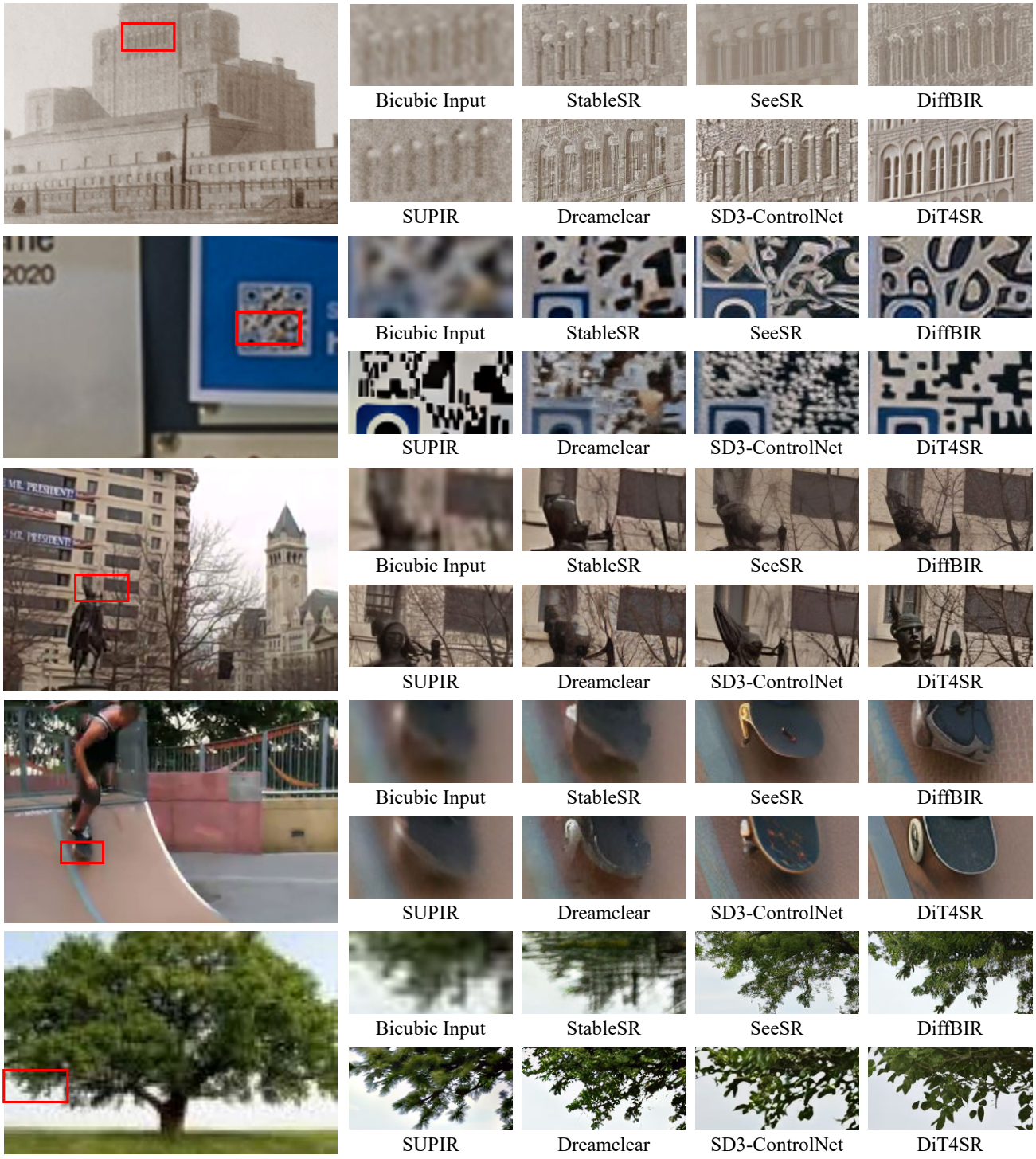


Figure 10. More qualitative comparisons with state-of-the-art Real-ISR methods on RealLQ250 (the first two rows) and RealLR200 (the last three rows). Our DiT4SR achieves the best performance in terms of image realism and detail generation while maintaining fidelity to the input LR image.

References

[1] 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. In *NeurIPS*, 2025. 1, 7

[2] blackforestlabs.ai. Flux, offering state-of-the-art performance image generation, 2024. Accessed: 2024-10-07. 1

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 6

[4] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 1

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1

[6] 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. 1, 7

[7] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*, 2020. 6

[8] 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. 7

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 1

[10] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 2, 3

[11] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[12] 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. 7

[13] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 7

[14] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024. 7

[15] 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. 7

[16] 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.

[17] 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*, 2024. 1, 6, 7

[18] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2024. 7

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023. 7