# Supplementary Material to "SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution"

Rongyuan Wu[1,2], Tao Yang[3], Lingchen Sun[1,2], Zhengqiang Zhang[1,2], Shuai Li[1,2] Lei Zhang[1,2,*]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute [3]ByteDance Inc.
{rong-yuan.wu, ling-chen.sun, zhengqiang.zhang, novak.li}@connect.polyu.hk
yangtao9009@gmail.com, cslzhang@comp.polyu.edu.hk

In this supplementary file, we provide the following materials:

- Ablation studies on the proposed LRE strategy and DAPE module;
- Complexity analysis;
- More real-world visual comparisons under scaling factor $4\times$

## 1. Ablation Study

We first discuss the effectiveness of the proposed LRE strategy. Then, we discuss the effectiveness of the proposed DAPE module, including its tagging capability and the roles of hard and soft prompts.

**Effectiveness of LRE.** We first show the Real-ISR performance of our SeeSR model on the *DIV2K-Val* and *DrealSR* datasets with and without the LRE strategy. The results are shown in Table 1. One can see that the LRE strategy improves the reference-based metrics, including both fidelity and perception based ones, while it weakens the no-reference metrics such as CLIPIQA. This is because the LRE strategy reduces the model's tendency to generate additional (but maybe unfaithful) textures by narrowing the gap between training and testing (see discussions in Section 3.4 of the main paper). Such an over-generation ability can be favorable by metrics like CLIPIQA, but they will introduce visually unpleasant artifacts, as shown in Fig. 3 of the main paper.

**Tagging Performance of DAPE.** In Table 2, we present the tagging performance of our DAPE module on the degraded images of *COCO-val* benchmark [2] based on three metrics: overall precision (OP), overall recall (OR), and average precision (AP). AP is the averaged precision calculated on different recall rates, which is similar to the detection metric. OP and OR are defined as:

$$\mathrm{OP} = \frac{\sum_i N_i^t}{\sum_i N_i^p}, \quad \mathrm{OR} = \frac{\sum_i N_i^t}{\sum_i N_i^g}, \tag{1}$$

where $C$ is the number of classes, $N_i^p$ is the number of images predicted for label $i$, $N_i^t$ is the number of images correctly predicted for label $i$, and $N_i^g$ is the number of ground truth images for label $i$.

We evaluate RAM [11] and DAPE with the default threshold. DAPE surpasses RAM in terms of OP and AP by 0.1 and 10.7, respectively. It also maintains superiority in OR, indicating that DAPE achieves significant improvements in tagging accuracy for degraded images. This improvement assists the T2I model in generating semantically accurate details when performing the Real-ISR task.

**Effectiveness of DAPE and Hard/Soft Prompts for Real-ISR.** DAPE improves the model's tagging performance on degraded images and consequently enhances the Real-ISR capability. To investigate the effectiveness of DAPE and the roles of its hard/soft prompts, we conducted the following four experiments in Real-ISR tasks.

1. We retrain SeeSR by removing the DAPE and RCA modules, which can be considered as applying ControlNet [10] directly to the Real-ISR task.

*Corresponding author. This work is supported by the Hong Kong RGC RIF grant (R5001-18) and the PolyU-OPPO Joint Innovation Lab.

Table 1. The Real-ISR performance of our SeeSR model with and without LRE on *DIV2K-Val* and *DrealSR* [7] benchmarks.

| Metrics | *DIV2K-Val* | | *DrealSR* | |
|---|---|---|---|---|
| | w/o LRE | w/ LRE | w/o LRE | w/ LRE |
| PSNR ↑ | 20.58 | 21.04 | 26.55 | 27.90 |
| LPIPS ↓ | 0.3942 | 0.3876 | 0.3952 | 0.3299 |
| FID ↓ | 32.53 | 32.79 | 158.04 | 151.88 |
| CLIPIQA ↑ | 0.7314 | 0.6834 | 0.7248 | 0.6708 |

Table 2. Comparison between RAM and DAPE on degraded images of *COCO-val* benchmark [2] for the tagging task.

| | OP ↑ | OR ↑ | AP ↑ |
|---|---|---|---|
| RAM [11] | 0.7929 | 0.3711 | 52.3 |
| DAPE | **0.8940** | **0.3751** | **63.0** |

Table 3. Ablation studies of DAPE on *DIV2K-Val* and *DrealSR* [7] benchmarks for the Real-ISR task.

| Exp | | (1) | (2) | (3) | (4) | SeeSR |
|---|---|---|---|---|---|---|
| Prompt Extractor | RAM [11] | ✗ | ✓ | ✗ | ✗ | ✗ |
| | DAPE | ✗ | ✗ | ✓ | ✓ | ✓ |
| Prompt Format | Hard Prompt | ✗ | ✓ | ✓ | ✗ | ✓ |
| | Soft Prompt | ✗ | ✓ | ✗ | ✓ | ✓ |
| *DIV2K-Val* | PSNR ↑ | 20.96 | 21.15 | 20.91 | 21.19 | 21.04 |
| | LPIPS ↓ | 0.4236 | 0.4156 | 0.4289 | 0.3859 | 0.3876 |
| | FID ↓ | 37.35 | 46.34 | 38.92 | 38.77 | 32.79 |
| | CLIPIQA ↑ | 0.6343 | 0.6097 | 0.6471 | 0.6751 | 0.6834 |
| *DrealSR* | PSNR ↑ | 27.64 | 27.31 | 27.45 | 28.14 | 27.90 |
| | LPIPS ↓ | 0.3130 | 0.3272 | 0.3285 | 0.3174 | 0.3299 |
| | FID ↓ | 176.26 | 161.69 | 164.57 | 157.63 | 151.88 |
| | CLIPIQA ↑ | 0.5693 | 0.6436 | 0.6410 | 0.6431 | 0.6708 |

2. We replace DAPE with RAM [11] and retrain the model.
3. During the inference of SeeSR, we provide only the hard prompts (*i.e.*, the tag) generated by DAPE to the text encoder of the T2I model.
4. During the inference of SeeSR, we provide only the soft prompts (*i.e.*, the representation embedding features) generated by DAPE to the T2I model.

The results of the four experiments are shown in Table 3. Moreover, the visual comparisons are shown in Fig. 1. From Table 3 and Fig. 1, we can have the following conclusions.

First, directly applying ControlNet to the Real-ISR task cannot achieve satisfactory results. Second, replacing DAPE with the original RAM would lead to a decrease in all perceptual metrics (*e.g.*, LPIPS and CLIPIQA). The semantics of the image content may also be changed (see Fig. 1). This is because the original RAM may generate inaccurate prompts (*e.g.*, the tag 'broccoli') from the degraded image. Third, the soft prompts work better in improving the numerical indices than the hard prompts, as well as sharper images. However, without hard prompts, the image semantics can be damaged, as can be seen from the lemons in Exp. (4) of Fig. 1. Finally, with both the hard and soft prompts in DAPE, perceptually realistic and semantically correct Real-ISR outputs can be produced.

## 2. Complexity Analysis

Table 4 compares the number of parameters of different Real-ISR models and their inference time to synthesize a $512 \times 512$ image from $128 \times 128$ input. All tests are conducted on one NVIDIA Tesla 32G-V100 GPU. We can have the following observations.

First, the GAN-based Real-ESRGAN and FeMaSR have much less model parameters and much faster inference speed

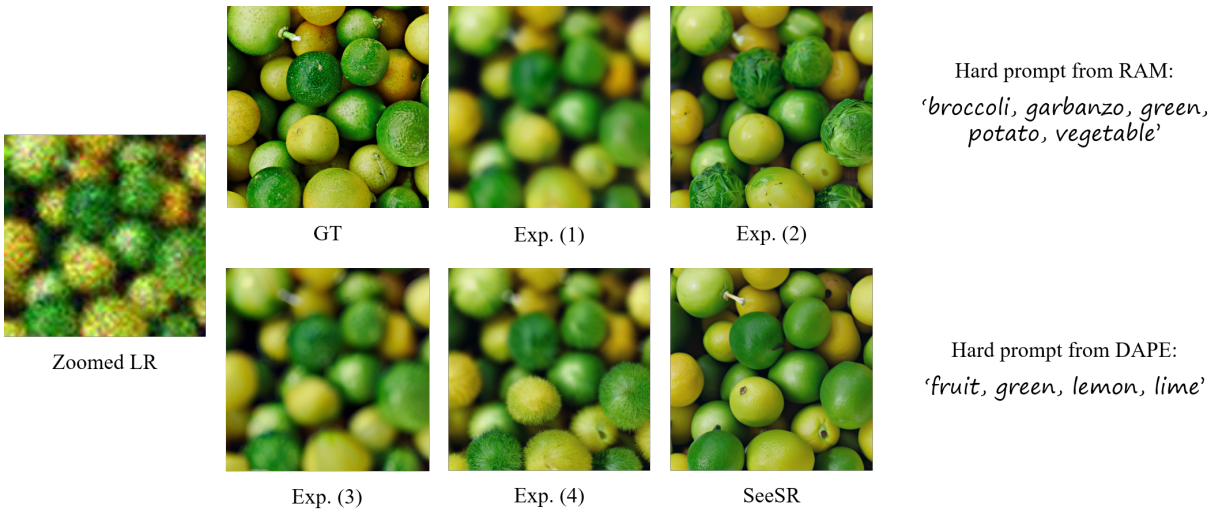


Figure 1. Visual comparison for the ablation study on DAPE. Exp. (1) directly applies ControlNet to perform Real-ISR, leading to blurry results. Exp. (2) replaces DAPE with RAM for generating prompts, which can produce sharper but semantically incorrect details. Exp. (3) applies hard prompts only and generates blurry results. Exp. (4) applies soft prompts only and exhibits semantic errors in details generation. With both hard and soft prompts in DAPE, SeeSR produces clear and semantically correct outputs.

Table 4. Complexity comparison between different methods. All the tests are conducted on one NVIDIA Tesla 32G-V100 GPU to synthesize $512 \times 512$ images from $128 \times 128$ inputs.

| Methods | Params | Inference Time-steps | Inference Time |
|---|---|---|---|
| Real-ESRGAN [6] | 16.7M | 1 | 0.09s |
| FeMaSR [1] | 28.3M | 1 | 0.12s |
| LDM [4] | 169.0M | 200 | 5.21s |
| StableSR [5] | 1409.1M | 200 | 18.70s |
| ResShift [9] | 173.9M | 15 | 1.12s |
| PASD [8] | 1900.4M | 20 | 6.07s |
| DiffBIR [3] | 1716.7M | 50 | 5.85s |
| SeeSR | 2283.7M | 50 | 7.24s |

than DM-based methods. Second, among the DM-based models, LDM and ResShift are much smaller than others because they employ relatively lightweight diffusion models. ResShift runs faster than LDM because it samples only 15 steps while LDM samples 200 steps. Third, StableSR, PASD, DiffBIR and our SeeSR are all based on the pre-trained T2I model. SeeSR has more parameters because it has a DAPE module (about 300M) finetuned from the RAM model. In terms of inference speed, PASD, DiffBIR and SeeSR are comparable, while StableSR is the slowest one because it samples 200 steps.

## 3. More Visualization Comparisons

We provide additional qualitative comparisons on real-world images. As shown in Fig. 2, SeeSR can generate sharper edges (case 2) and semantically faithful details (the window railing in case 1, the teeth in case 3, and the vein textures in case 4). Other methods are either blurry or produce unpleasant artifacts.

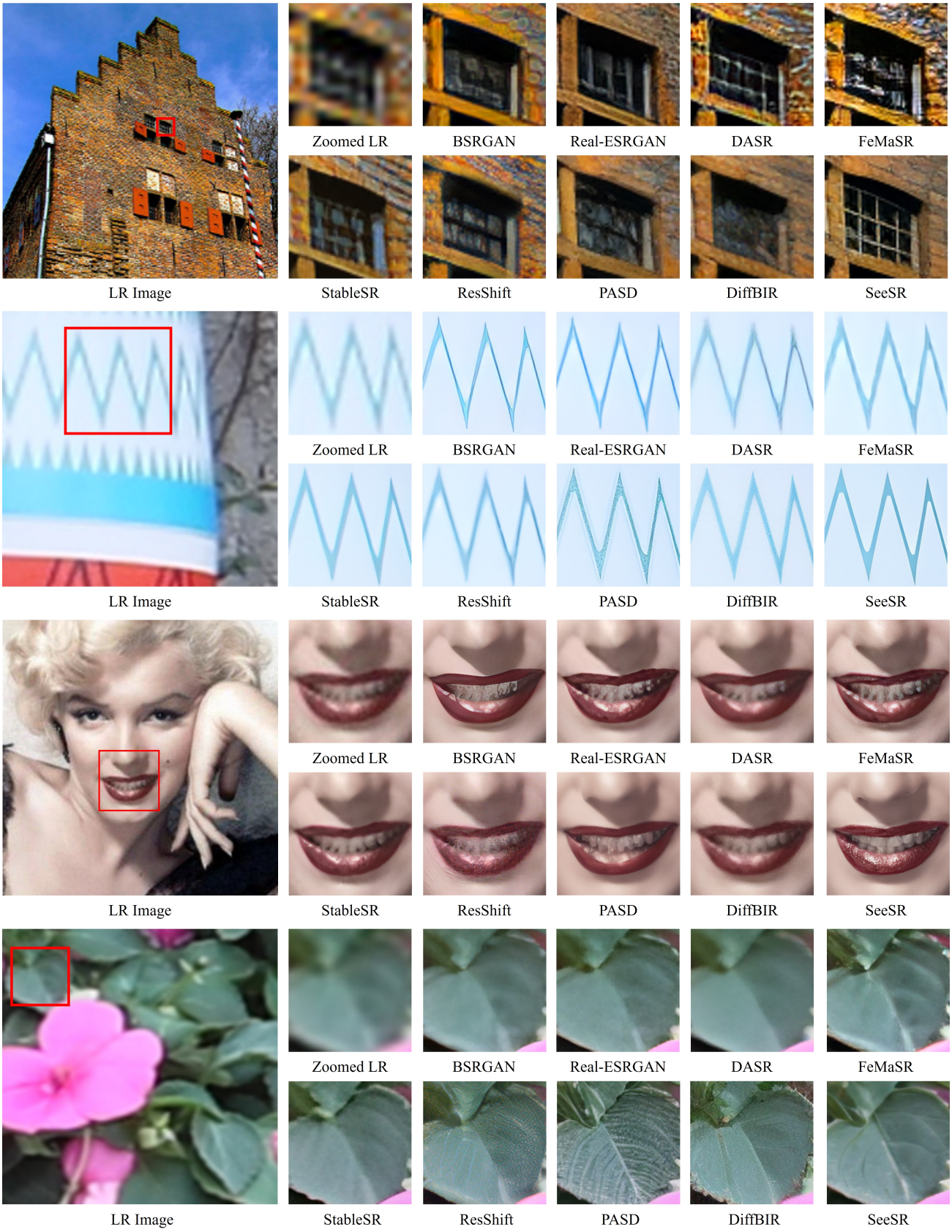


Figure 2. Qualitative comparisons of different methods on real-world examples. Please zoom in for a better view.

## References


[1] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1329–1338, 2022. 3

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1, 2

[3] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 3

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[5] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3

[6] 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. 3

[7] 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

[8] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 3

[9] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *arXiv preprint arXiv:2307.12348*, 2023. 3

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[11] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 1, 2