XPSR: Cross-modal Priors for Diffusion-based Image Super-Resolution

—— Supplementary Material ——

Yunpeng Qu[1,2†], Kun Yuan[2†‡], Kai Zhao[2], Qizhi Xie[1,2], Jinhua Hao[2], Ming Sun[2] and Chao Zhou[2]

[1] Tsinghua University, China, Beijing
[2] Kuaishou Technology, China, Beijing
{qyp21,xqz20}@mails.tsinghua.edu.cn
{yuankun03,zhaokai05,haojinhua,sunming03,zhouchao}@kuaishou.com

1 Appendix

1.1 Implementation Details of XPSR

Image Encoder Stable Diffusion [4] utilizes a VAE Encoder to map the original 512×512 image to a latent representation of 64×64 dimensions, enabling iterative denoising in the latent space. To align with the scale of Stable Diffusion, the LR images are also upsampled by a factor of 4 to match the same resolution. Then, through a pyramid-like image encoder [13], the scale is gradually reduced to a 64×64 feature space vector. The image encoder consists of three layers, with each layer comprising two convolutional layers where the stride of the second convolution is 2×2, resulting in a halving of the feature map size.

In our pixel-space constraint, we use linear layers to map the feature maps of i-th layer $\mathbf{X}_i$ to RGB images $\hat{x}_i \in \mathbb{R}^{\frac{512}{2^i} \times \frac{512}{2^i} \times 3}$, which corresponds to the i-th scale of the HR image after downsampling. The L_1 loss is utilized to ensure that the extracted features closely resemble the semantic content of the HR image.

ControlNet The ControlNet component, with a trainable copy of the Unet Encoder from Stable Diffusion, extracts multi-scale features through a pyramid structure. Within ControlNet, the features from LR images are further reduced from a dimension of 64×64 to a latent representation of 8×8. The multi-scale conditional controls extracted by ControlNet are connected to the Unet through zero-convolution residual connections and *conditional attention*. Therefore, the conditional features at scale j are also mapped to the channels of the latent representation through a linear transformation as $\hat{z}_j \in \mathbb{R}^{\frac{64}{2^j} \times \frac{64}{2^j} \times 4}$, corresponding to the i-th scale downsampled result of the HR latent. The L_1 loss is also applied to enforce the constraint in the latent space.

1.2 Additional Experimental Results

† Equal contribution.
‡ Project leader.

Table 1: Results of user study on real-world images.

Methods	BSRGAN	Real-ESRGAN	StableSR	PASD	SeeSR	**XPSR(Ours)**
Selection Rates	0.7%	0.9%	5.3%	14.0%	14.8%	**64.3%**

Table 2: Quantitative comparison with SOTA methods on real-world dataset with no reference images. Red and blue colors are the best and second-best performance.

Dataset	Metrics	GAN-based SR			Diffusion-based SR					
		BSRGAN	Real-ESRGAN	SwinIR	LDM	StableSR	DiffBIR	PASD	SeeSR	XPSR
RealLR200	MANIQA↑	0.3671	0.3633	0.3741	0.3049	0.3688	0.4288	0.4295	0.4844	0.5589
	CLIPIQA↑	0.5698	0.5409	0.5596	0.5253	0.5935	0.6452	0.6325	0.6553	0.7524
	MUSIQ↑	64.87	62.96	63.55	55.19	63.29	62.44	66.50	68.37	69.30

User Study To thoroughly evaluate the performance of our XPSR in real-world scenarios, we conduct a user study on 50 LR real-world images randomly sampled from *DrealSR* [1] and *RealSR* [8]. We compare our XPSR with five other ISR methods, including: BSRGAN [12], Real-ESRGAN [7], StableSR [6], PASD [11] and SeeSR [9]. For each image, the participants were simultaneously shown the LR image along with the restoration results from all ISR methods, and they were then instructed to select the best ISR result for the LR image. A total of 20 participants were invited to the user study and made a total of 20×50 votes, which are shown in Tab. 1. Our method achieved **the highest selection rate of 64.3%**, which is 4 times higher than the second-ranked method, showcasing the powerful application capabilities of XPSR in real-world scenarios.

Comparisons on real-world images To evaluate the capabilities of our method in in-the-wild scenarios, we conduct tests on the *RealLR200* dataset [9]. The *RealLR200* dataset consists of 200 real-world images, incorporating results collected from different studies [3, 7] as well as some images collected from the internet. Due to the lack of available reference HR images for these real-world images, we only utilize three non-reference IQA metrics, including MANIQA [10], MUSIQ [2], and CLIPIQA [5]. The quantitative results are shown in Tab. 2.

It can be observed that our XPSR performs **the best in all three metrics**, which is consistent with the results obtained on other datasets mentioned in the main text. In addition, we have visualized some results in Fig. 2, which indicate that XPSR is capable of recovering more realistic details compared to other methods, including lifelike facial features, intricate textures of the fur, and clearer tree leaves. The aforementioned results clearly demonstrate the powerful image restoration capabilities of XPSR even in in-the-wild scenarios.

The impact of MLLMs on ISR performance. XPSR relies on MLLMs to obtain low-level and high-level semantic embeddings. Therefore, it is necessary to explore how the limitations of MLLMs may impact the performance of the method. We have explored based on the following three aspects. **(1)** *Precision.* If an MLLM fails to understand LR, restoration results may be unrealistic

Table 3: Ablation on the effect of MLLMs.

Setting	*DrealSR*					*RealSR*				
	SSIM↑	LPIPS↓	FID↓	MANIQA↑	MUSIQ↑	SSIM↑	LPIPS↓	FID↓	MANIQA↑	MUSIQ↑
original	0.7220	0.3864	164.68	0.5713	67.84	0.6870	0.3517	141.95	0.6059	70.23
+ LLaVA-13b	0.7237	0.3870	165.70	0.5760	67.94	0.6920	0.3508	141.94	0.6099	70.32
+ Simple desc.	0.7190	0.3859	168.10	0.5697	67.74	0.6880	0.3537	148.93	0.5991	69.89
+ Downscale 64	0.7212	0.3884	167.80	0.5706	67.44	0.6856	0.3588	144.54	0.5966	69.86
+ Upscale 256	0.7204	0.3876	167.38	0.5733	67.89	0.6855	0.3525	142.45	0.6079	70.36
+ Gaussian blur	0.7121	0.3926	175.77	0.5730	67.16	0.6812	0.3624	146.62	0.5945	69.38
+ Jpeg comp.	0.7137	0.3902	170.83	0.5690	67.42	0.6822	0.3578	145.00	0.6040	69.70

Fig. 1: Limitations of diffusion-based methods. Due to their limited semantic understanding, the restored content may be unrelated to the original image.

or incorrect. Hence, we utilize the larger and more accurate LLaVA-13b model to generate the prompt. **(2)** *Completeness.* As shown in Sec.3.3 of the paper, XPSR employs the detailed descriptions generated by the MLLMs as guidance. For comparison, we build a simple prompt limited to 20 words, containing only object or distortion categories, to assess the impact of semantic completeness on image restoration. **(3)** *Robustness.* MLLM might struggle with diverse and complex image conditions, limiting its generalization ability. Therefore, we further degrade the LR images to obtain semantic cues under various complex degradation scenarios (*e.g.*, resolution, JPEG compression, Gaussian blur).

As given in Tab. 3, more precision and complete descriptions help to achieve better results, while the model suffers when LR owns severe degradations (*i.e.*, JPEG compression and Gaussian blur). The resolution has little impact on the final result, which means that MLLM can still obtain accurate semantic descriptions. **Notably, XPSR is orthogonal to MLLMs,** and better MLLMs in these three aspects can further advance XPSR.

1.3 Limitations

In this section, we discuss the limitations of diffusion-based methods in the ISR task. We have found that although diffusion models possess powerful generative capabilities and can produce realistic and detailed images, their understanding of specific scenes is limited. As a result, they can sometimes generate semantic-unrelated content. In Fig. 1, we present two examples.

For the left case, although our XPSR model recognizes the main subject of the original image as a cat, it overlooks the cartoon-style nature of the image,

resulting in the generated cartoon character exhibiting a realistic fur texture that is typically found only on real cats. The same issue occurs in other diffusion-based models, as demonstrated in the case on the right. Although SeeSR recognizes the person's motion, it fails to realize that the original image depicts the famous character Spider-Man, restoring a clear but unrelated portrait. Therefore, we consider that further enhancing the semantic understanding of different scenes is crucial for the successful application of diffusion models in ISR, which is also the motivation behind our XPSR. We strongly believe that this approach holds significant potential for exploration.

1.4 More Visual Results

In this section, we provide additional visualization results. Fig. 3 displays the high-level and low-level semantic prompts generated by LLaVA for different images, which align well with human perception. This illustrates the reliability of MLLM in incorporating cross-model semantic priors. In Fig. 4 and Fig. 5, we present additional comparative results with other methods, which further demonstrate the powerful capabilities of XPSR in generating high-fidelity and high-realistic images.

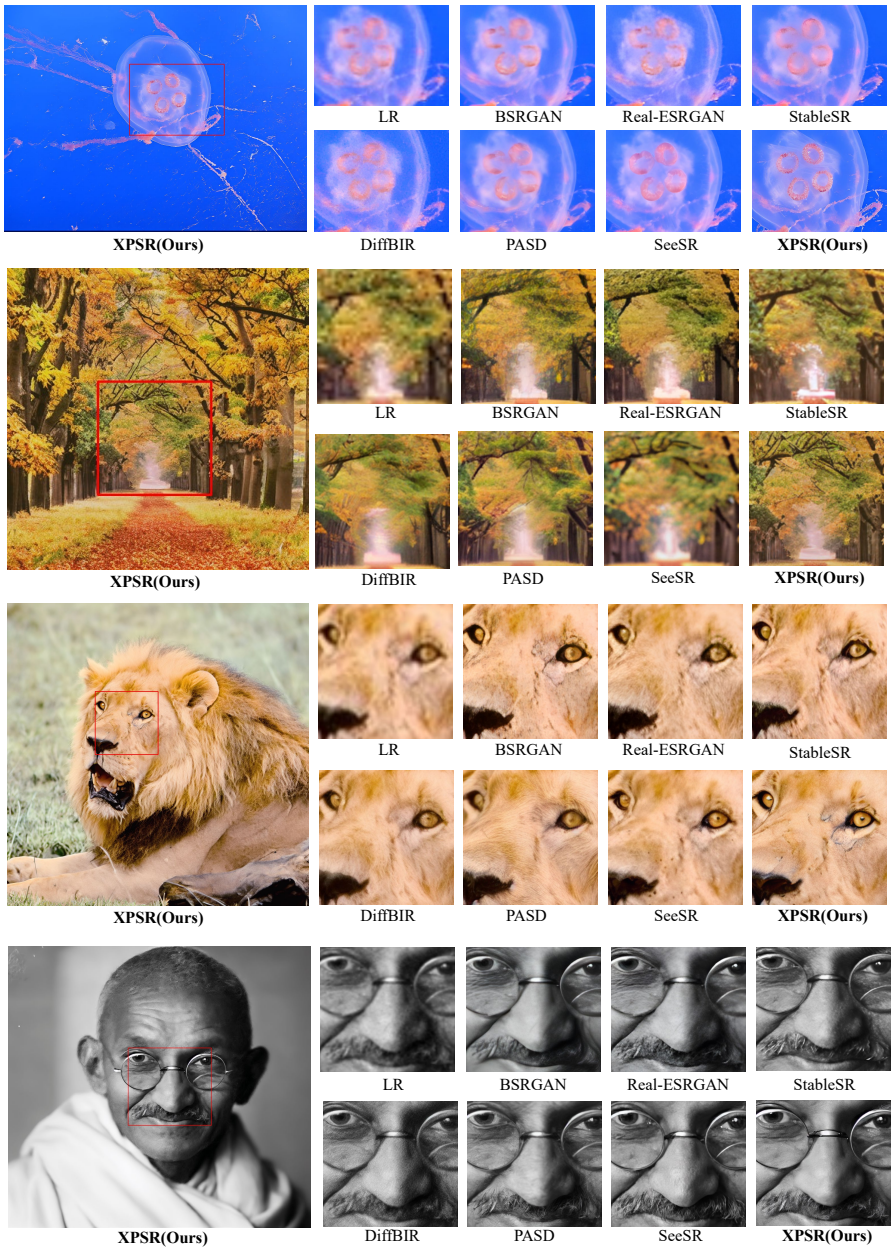

Fig. 2: Qualitative comparisons with different SOTA methods on real-world images. **Zoom in for a better view**.

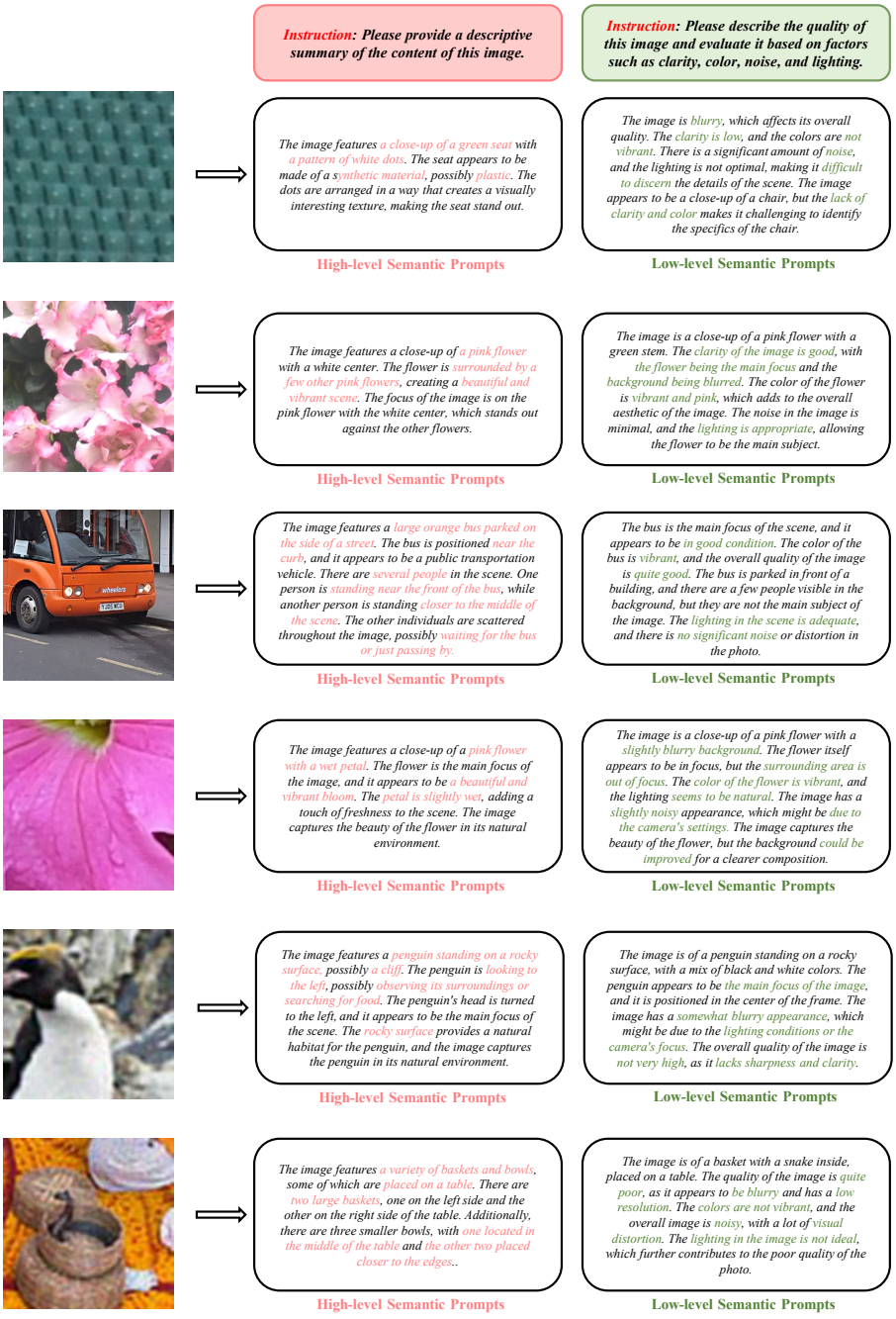

Fig. 3: Additional cases show that LLaVA can generate high- and low-level semantic prompts consistent with human perception for both high- and low-quality images.

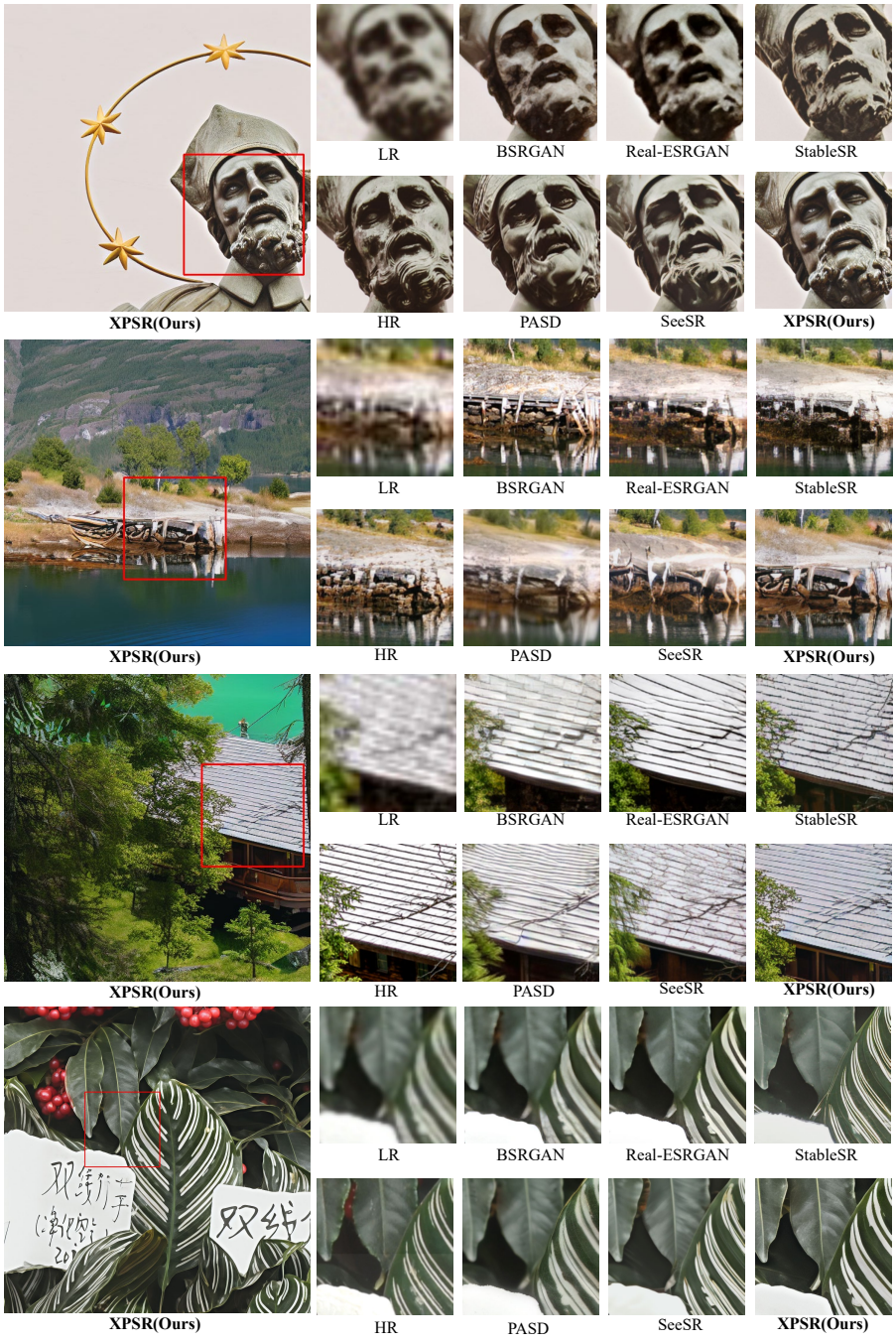

Fig. 4: Additional qualitative comparisons with different SOTA methods (Part 1). **Zoom in for a better view**.

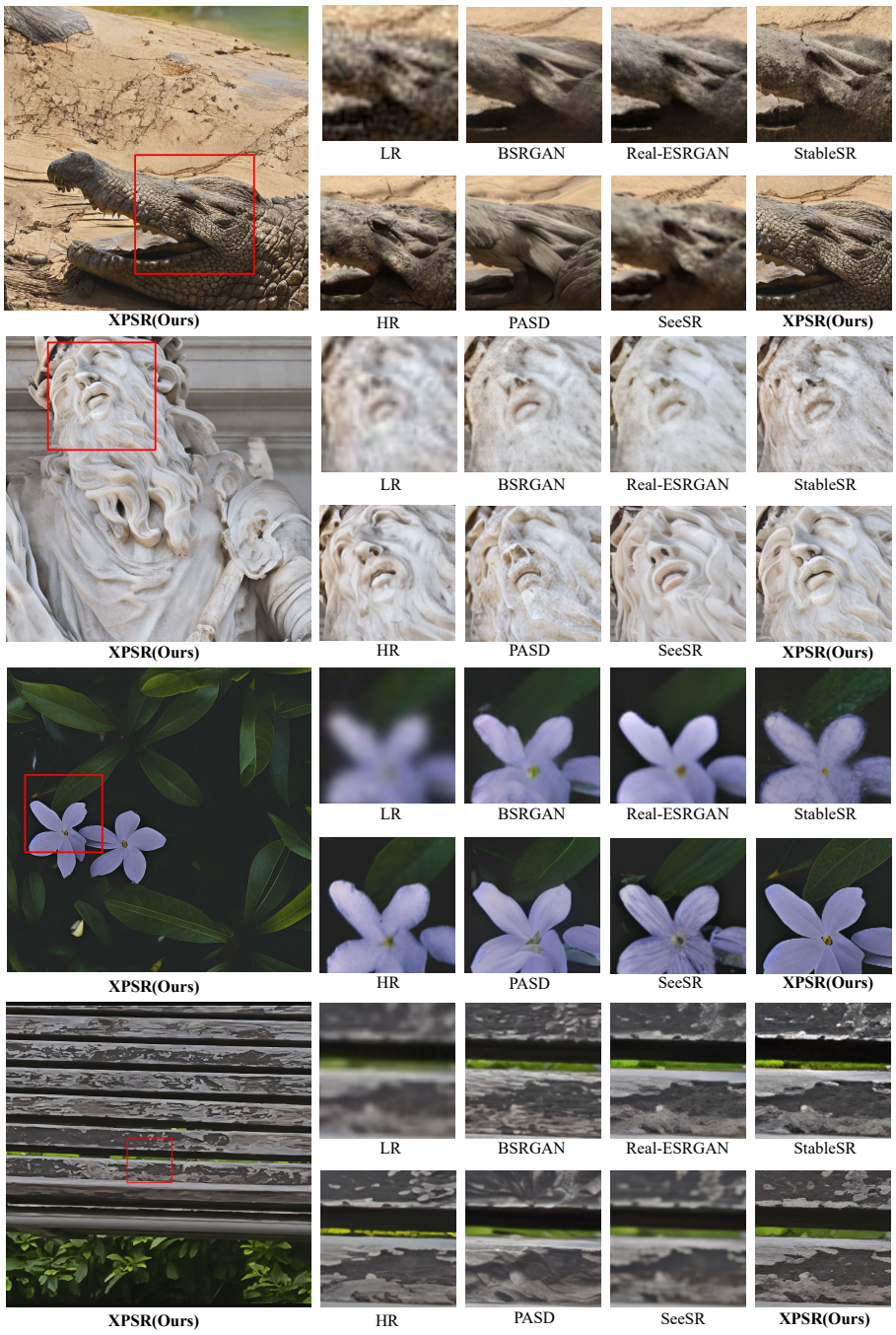

Fig. 5: Additional qualitative comparisons with different SOTA methods (Part 2). **Zoom in for a better view**.

References

1. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV. pp. 3086–3095. IEEE (2019)
2. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
3. Lin, X., He, J., Chen, Z., Lyu, Z., Fei, B., Dai, B., Ouyang, W., Qiao, Y., Dong, C.: Diffbir: Towards blind image restoration with generative diffusion prior. CoRR **abs/2308.15070** (2023)
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
5. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
6. Wang, J., Yue, Z., Zhou, S., Chan, K.C.K., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. CoRR **abs/2305.07015** (2023)
7. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCVW. pp. 1905–1914. IEEE (2021)
8. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: ECCV (8). Lecture Notes in Computer Science, vol. 12353, pp. 101–117. Springer (2020)
9. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. CoRR **abs/2311.16518** (2023)
10. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: multi-dimension attention network for no-reference image quality assessment. In: CVPR Workshops. pp. 1190–1199. IEEE (2022)
11. Yang, T., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. arXiv preprint arXiv:2308.14469 (2023)
12. Zhang, K., Liang, J., Gool, L.V., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV. pp. 4771–4780. IEEE (2021)
13. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)