FaithDiff: Unleashing Diffusion Priors for Faithful Image Super-resolution - Supplemental Material -

Junyang Chen Jinshan Pan Jiangxin Dong†
School of Computer Science and Engineering, Nanjing University of Science and Technology
https://jychen9811.github.io/FaithDiff_page/

Overview

In this supplemental material, we first present the detailed implementation of the proposed approach during training and inference in Section A. We provide more analysis and discussions of our method in Section B. More experimental results are shown in Section C.

†Corresponding author

A. Implementations of FaithDiff

We describe the training and inference processes of FaithDiff in Algorithm 1 and Algorithm 2.

Algorithm 1 Training of FaithDiff.

Input: α, β: learning rates of LQ encoder and diffusion model with alignment module
Input: I^{HQ}, I^{LQ}, C^{Text}: high-quality image, low-quality image and text description
Input: $\hat{\epsilon}_\theta$, θ_{lq}, θ_{ec}, θ_{text}: latent diffusion model with alignment module, LQ encoder, VAE encoder and text encoder

while not converged **do**
  Sample a batch of image data and text descriptions in $\{I^{HQ}, I^{LQ}, C^{Text}\}$;
  Sample time step t from $Uniform(1, ..., T)$ and noise ϵ from $\mathcal{N}(0, \mathbf{I})$;
  Extract f^{HQ} from I^{HQ} via θ_{ec};
  Extract f^{LQ} from I^{LQ} via θ_{LQ};
  Extract c from C^{Text} via θ_{text};
  Add noise to f^{HQ} to fetch x_t^{HQ};
  Compute $L(\epsilon, \hat{\epsilon}_\theta(x_t^{HQ}, f^{LQ}, c, t))$;
  Update parameters of θ_{lq} and $\hat{\epsilon}_\theta$ with gradient descent:
  $\theta_{lq}^{update} \leftarrow \theta_{lq} - \alpha \nabla_{\theta_{lq}} L(\epsilon, \hat{\epsilon}_\theta(x_t^{HQ}, f^{LQ}, c, t))$;
  $\hat{\epsilon}_\theta^{update} \leftarrow \hat{\epsilon}_\theta - \beta \nabla_{\hat{\epsilon}_\theta} L(\epsilon, \hat{\epsilon}_\theta(x_t^{HQ}, f^{LQ}, c, t))$;
end while

Output:
  Updated parameter θ_{lq}^{update} and $\hat{\epsilon}_\theta^{update}$.

Algorithm 2 Inference of FaithDiff.

Input: I^{LQ}, C^{Text}: low-quality image and text description
Input: $\hat{\epsilon}_\theta$, θ_{lq}, θ_{dc}, θ_{text}: latent diffusion model with alignment module, LQ encoder, VAE decoder and text encoder

Sample noisy latent x_T^{HQ} from $\mathcal{N}(0, \mathbf{I})$;
Extract f^{LQ} from I^{LQ} via θ_{LQ};
Extract c from C^{Text} via θ_{text};
for $t = T, ..., 1$ **do**
  Sample z from $\mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $z = 0$;
  Predict the noisy latent at time step $t - 1$:
  $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t^{HQ} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_\theta(f_t^a, c, t)) + \sigma_t z$;
end for
Generate restored images:
$I^{Res} = \theta_{dc}(x_0^{HQ})$.

Output:
  Restored image I^{Res}

B. In-Depth Analysis

In this section, we provide additional ablation studies and analysis on the proposed FaithDiff.

B.1. Efficiency of FaithDiff

In the main paper, we compare the run-time performance of diffusion-based SR methods [1, 3, 8–10] in Table 4. Note that existing diffusion-based SR methods rely on ControlNet [12] to steer the diffusion process with the provided LQ image. In contrast, benefiting from the proposed unified feature optimization strategy, our FaithDiff does not need ControlNet and is able to adopt a simple yet effective alignment module to guide the diffusion process. To further demonstrate the efficiency of FaithDiff, for a fair comparison, we compare with the baseline 'FT EN & Fix DM' in Table 6 of the main paper, which uses ControlNet with the SFT layer [10], in terms of GPU memory consumption and inference time. The comparison results in Table 7 show that our approach requires about 2.55 seconds on images of 1024×1024 pixels, which is 2.31 times faster than the ControlNet-based baseline. In addition, our approach reduces the GPU memory consumption by approximately 37% and 65% for the training and inference stages, respectively. Thus, our FaithDiff is more efficient.

Table 7. GPU memory consumption and inference time. #GPU Mem. denotes the maximum of GPU memory consumption, evaluated using FP16 precision.

Method	Training			Inference		
	Batch Size	DeepSpeed Stage	#GPU Mem. (M)	Inference Step	Inference Time (s)	#GPU Mem. (M)
Ours	10	2	50,990	20	2.55	9,301
FT EN & Fix DM	10	2	80,628	20	5.89	26,510

B.2. Comparison with ControlNet-based variants

We have compared against ControlNet in Tab. 6 (ours *vs.* 'FT EN & Fix DM'). We further compare with ControlNet-XS in Table 8, where our method performs better. The inference steps of all methods are set as 20.

Table 8. Quantitative comparison with different ControlNet-based variants.

Dataset	DIV2K-Val		RealPhoto60		Running
Metric	LPIPS ↓	MUSIQ ↑	MUSIQ ↑	CLIPIQA+ ↑	Time (s)
ControlNet-XS	0.3779	63.02	67.59	0.6235	2.61
ControlNet	0.3370	66.03	69.66	0.6412	4.62
Ours	0.3080	66.28	72.74	0.6527	2.55

B.3. More details and quantitative results on the RealDeg Dataset

In Section 4 of the main paper, to evaluate the performance of our method in real-world scenarios, we collect a dataset of 238 images with unknown degradations, consisting of old photographs, social media images, and classic film stills. Old photographs include black-and-white ones and faded-color ones. For social media images, we first collect images from Unsplash and then upload them to various social media platforms, undergoing one or multiple rounds of cross-platform processing. Classic film stills are selected from films spanning from 1980s to 2000s. The RealDeg dataset contains diverse categories of content including buildings, animals, and various natural elements. In addition, the image resolution is at least 720×720 pixels. Some examples are shown in Figure 7. We further evaluate the proposed method on the RealDeg datasets in Table 9.

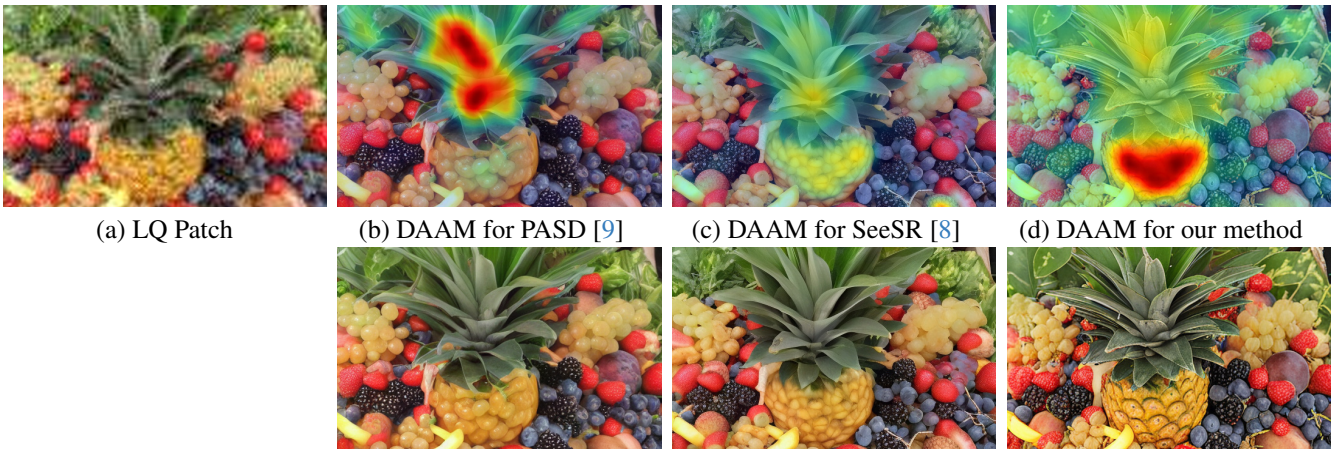
Table 9. Quantitative comparison with state-of-the-art methods on RealDeg dataset. The best and second performances are marked in red and blue, respectively.

Benchmarks	Metrics	Real-ESRGAN [7]	BSRGAN [11]	StabeSR [6]	DiffBIR [3]	PASD [9]	SeeSR [8]	DreamClear [1]	SUPIR [10]	Ours
RealDeg	MUSIQ ↑	52.64	52.08	53.53	58.22	47.31	60.10	56.67	51.50	61.24
	CLIPIQA+ ↑	0.3396	0.3520	0.3669	0.4258	0.3137	0.4315	0.4105	0.3468	0.4327
	PaQ-2-PiQ ↑	70.29	70.20	69.54	70.97	66.07	71.55	70.88	69.87	72.41
	NIQE ↓	3.8825	4.7553	4.6347	4.2155	4.9009	4.4827	3.7912	3.9643	3.9001
	MANIQA ↑	0.5004	0.4895	0.4986	0.5487	0.4571	0.5241	0.5335	0.5248	0.5703

Figure 7. Examples from the RealDeg dataset.

B.4. More visualization results of DAAMs

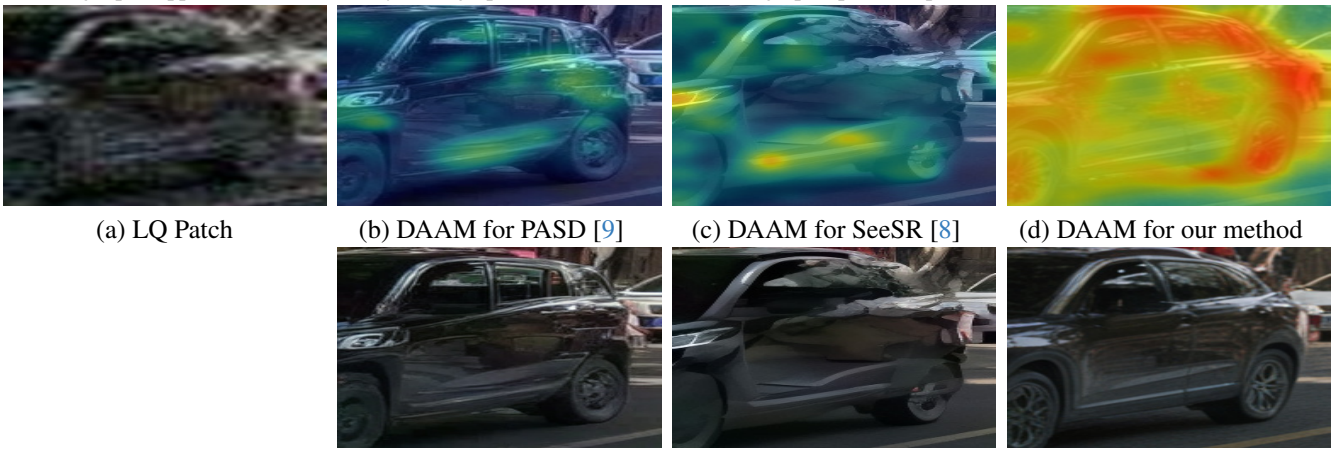
We present more diffusion attentive attribution maps (DAAMs) [4] for different scenes. Figures 8-10 show that when existing diffusion-based methods struggle to extract accurate structural information from LQ images, the text embeddings may generate low responses in the LQ features, hindering the ability of diffusion priors to restore faithful structures. In contrast, our method can explore more useful information from LQ images and produce a more plausible DAAM [4].



(a) LQ Patch (b) DAAM for PASD [9] (c) DAAM for SeeSR [8] (d) DAAM for our method

(e) PASD [9] (f) SeeSR [8] (g) Ours

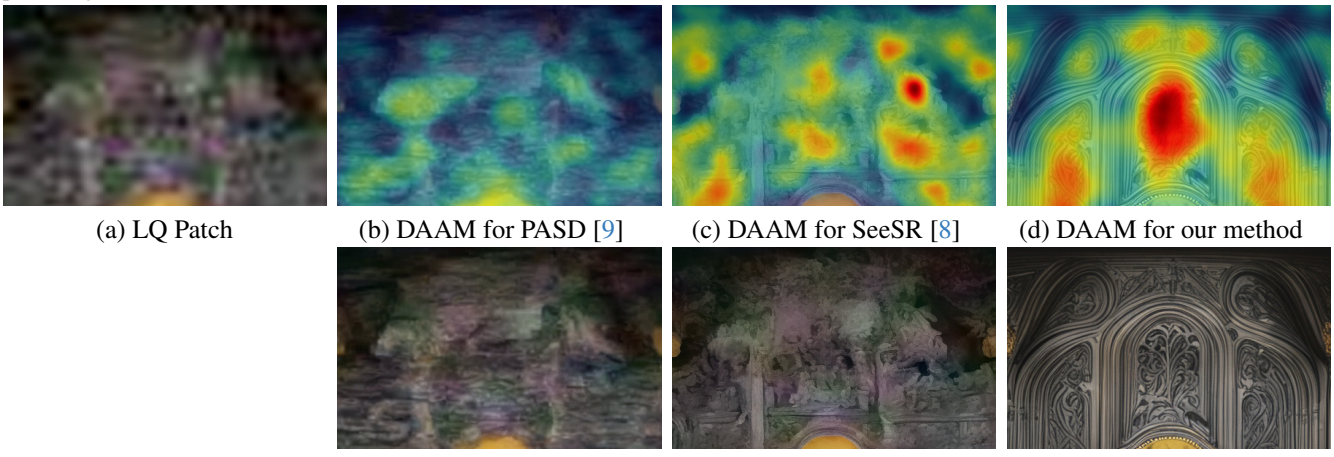
Figure 8. Visualization of DAAMs [4] for 'pineapple' from 'An image of a colorful fruit basket surrounded by various fruits and nuts, including a pineapple, red strawberries, yellow grapes, black blackberries, blue grapes, peaches, plums, and walnuts'.



(a) LQ Patch (b) DAAM for PASD [9] (c) DAAM for SeeSR [8] (d) DAAM for our method

(e) PASD [9] (f) SeeSR [8] (g) Ours

Figure 9. The DAAM [4] for 'cars' from 'The image depicts a city street with a man on a scooter, cars parked along the sides, and trees providing shade'.



(a) LQ Patch (b) DAAM for PASD [9] (c) DAAM for SeeSR [8] (d) DAAM for our method

(e) PASD [9] (f) SeeSR [8] (g) Ours

Figure 10. The DAAM [4] for 'carvings' from 'The image shows a yellow upholstered chair on an ornate dark wooden platform, surrounded by intricate carvings, giving a classic and grand atmosphere'.

B.5. Effects of inference steps and classifier free guidance (CFG)

We evaluate the effect of inference steps and classifier free guidance (CFG) by varying their values from 10 to 50 and 2 to 8, respectively, in Tab. 10. We empirically use 20 inference steps and set CFG scale as 5 as a trade-off between quality and fidelity.

Table 10. Quantiative performance (MUSIQ / LPIPS) on the dataset of DIV2K-Val for various CFG and inference steps.

CFG \ Steps	10	20	30	40	50
2	63.82 / 0.2954	63.80 / 0.2896	63.35 / 0.2958	63.26 / 0.2983	63.15 / 0.3011
5	66.71 / 0.3127	66.28 / 0.3080	65.95 / 0.3120	65.83 / 0.3139	65.72 / 0.3158
8	67.62 / 0.3334	67.05 / 0.3328	66.78 / 0.3365	66.64 / 0.3385	66.56 / 0.3404

B.6. Effect of the inference strategy on FaithDiff

We provide an ablation study on inference strategy [9] for PASD, SeeSR, and FaithDiff. Adopting fixed LDM may generate features unrelated to the LQ features with random noisy inputs, so inference strategies are used to alleviate this problem. In contrast, we align the LQ features with the noisy input by the proposed alignment module and unleash the LDM to explore useful information and boost faithful image SR. In this way, we minimize the negative effect of random noisy inputs and inference strategies have little impact on our method (see Tab. 11).

Table 11. Differences between with and without inference strategy (△ =w / IS - w/o IS) on synthetic datasets *i.e.*, DIV2k-Val, where IS denotes inference strategy.

Method	PASD		SeeSR		Ours	
Metrics	△PSNR	△MUSIQ	△PSNR	△MUSIQ	△PSNR	△MUSIQ
DIV2K	0.78	-4.85	0.35	-0.74	0.07	-0.03
LSDIR	0.58	-3.99	0.21	-0.46	0.04	-0.04

B.7. FaithDiff on SD 2-1

As shown in Table 12, our FaithDiff outperforms all competing methods using the SD 2-1 backbone, achieving improvements of at least 1.55 in MUSIQ [2] and 0.09 in CLIPIQA+ [5] on the RealPhoto60 [10] benchmark.

Table 12. Quantitative comparison with state-of-the-art methods on real-world benchmarks. The best and second performances are marked in red and blue, respectively.

Benchmarks	Metrics	Real-ESRGAN [7]	BSRGAN [11]	StabeSR [6]	DiffBIR [3]	PASD [9]	SeeSR [8]	DreamClear [1]	SUPIR [10]	Ours
RealPhoto60 [10]	MUSIQ ↑	59.29	45.46	57.89	63.67	64.53	70.80	70.46	70.26	72.35
	CLIPIQA+ ↑	0.4389	0.3397	0.4214	0.4935	0.4786	0.5691	0.5273	0.5528	0.6591

C. Quantitative Comparisons

In this section, we first present more visual comparisons with state-of-the-art methods [7, 9, 11] on synthetic images with mild, medium and severe degradation effects in Figure 11 and Figure 12, where our method can generate faithful structures (*e.g.*, stripes in the first and third examples in Figure 11) and realistic details (*e.g.*, grass in the second example and textures of the butterfly in the fourth example in Figure 11).

Then, we present additional visual comparisons with state-of-the-art methods [1, 3, 7–10] on real-world benchmarks. As shown in Figures 13-22, our proposed method can recover more faithful structural details.

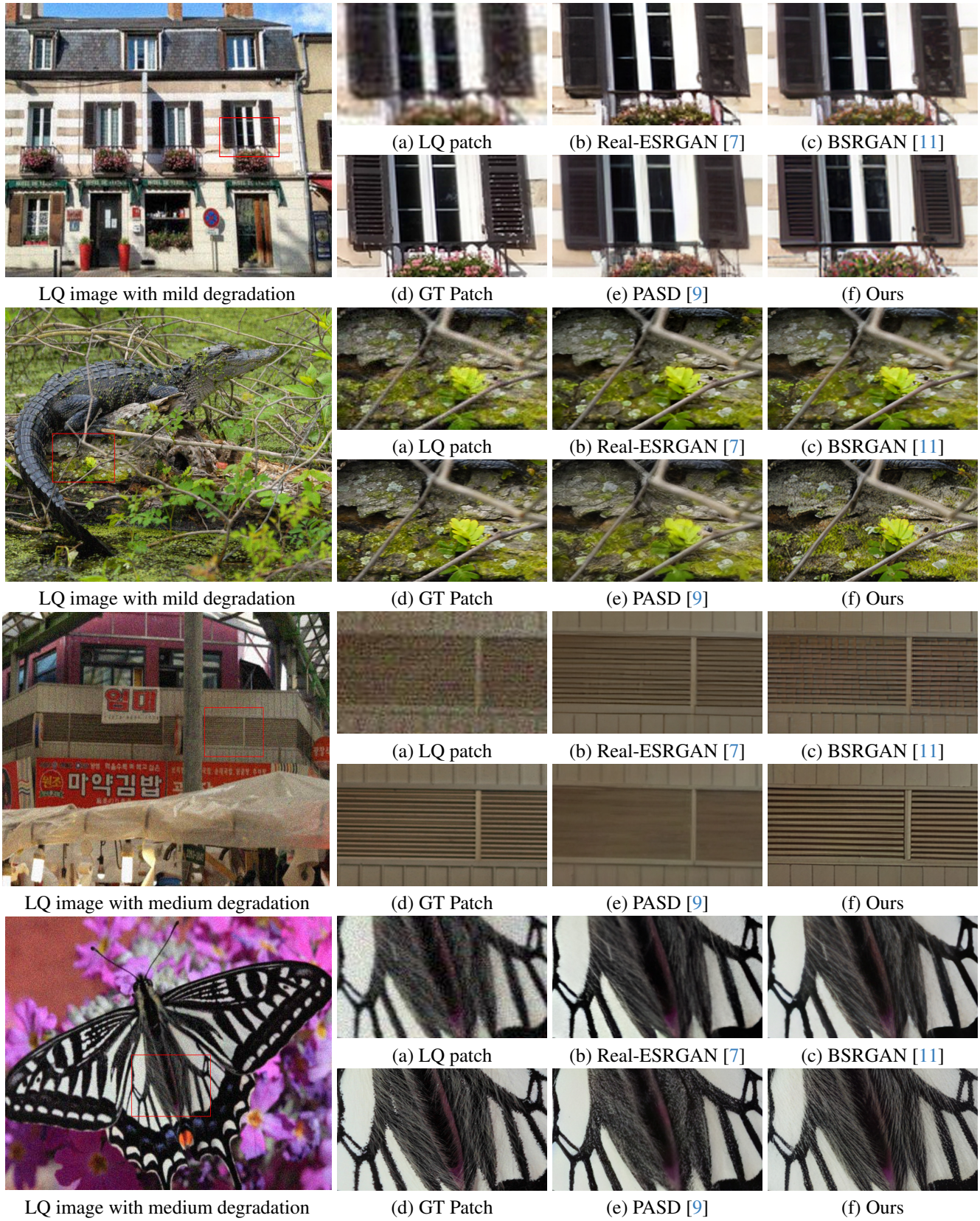


Figure 11. Image SR results on examples from the synthetic datasets. The proposed method recovers much clearer structural details in (f).

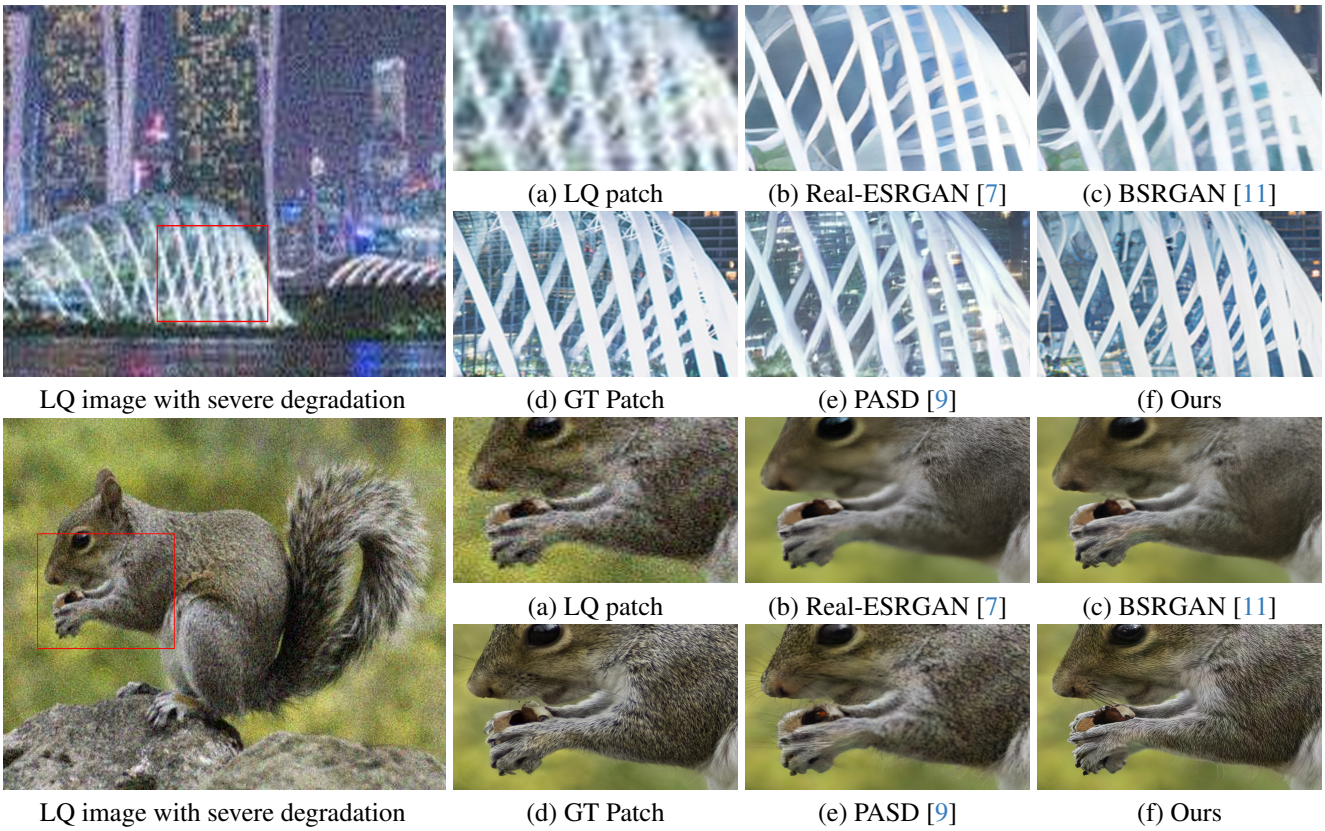
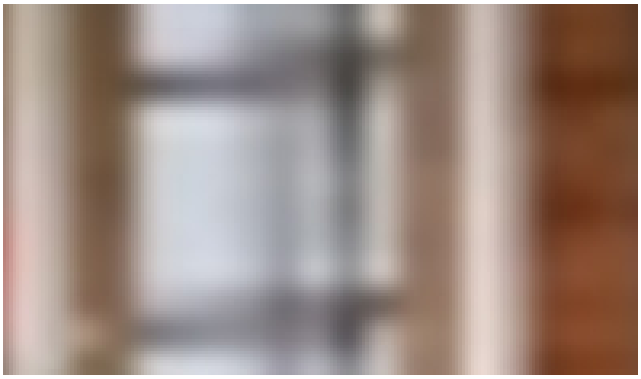


Figure 12. Image SR results on examples from the synthetic datasets. The proposed method recovers much clearer structural details in (f).



(a) LQ Patch

(b) Real-ESRGAN [7]

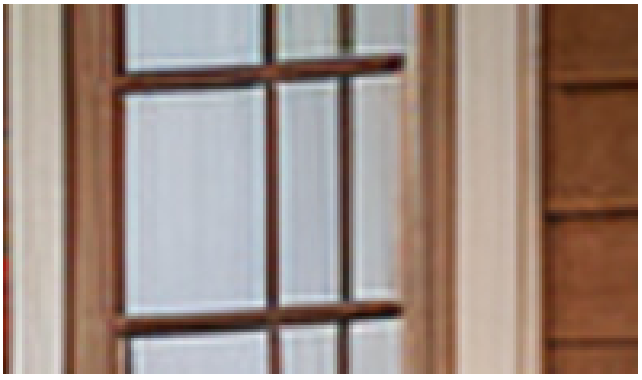
(c) PASD [9]

(d) DiffBIR [3]

(e) SeeSR [8]

(f) DreamClear [1]

(g) SUPIR [10]



(h) Ours

Figure 13. Image SR results on an example from the RealPhoto60 [10] dataset. Compared to competing methods, our proposed method recovers more realistic details (*e.g.*, the strips of window in (h)).

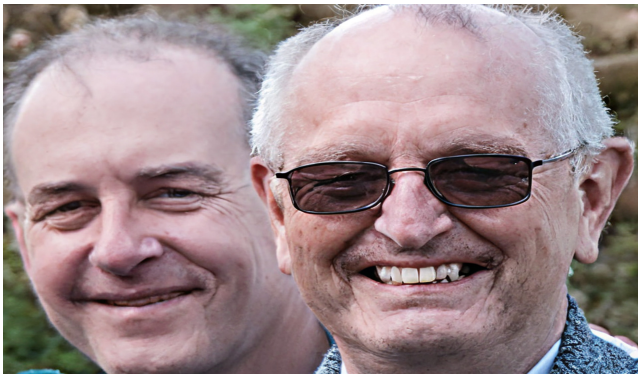
(a) LQ Patch

(b) Real-ESRGAN [7]

(c) PASD [9]

(d) DiffBIR [3]

(e) SeeSR [8]



(f) DreamClear [1]

(g) SUPIR [10]

(h) Ours

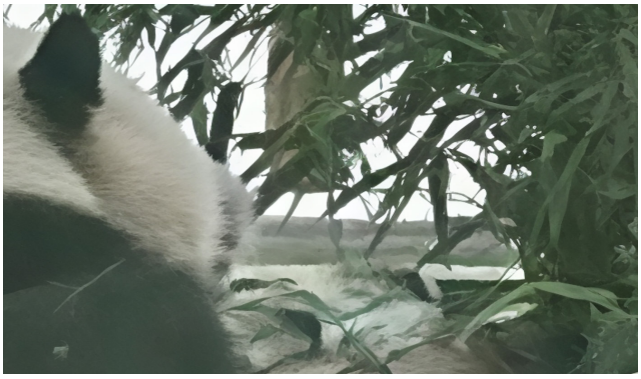
Figure 14. Image SR results on an example from the RealPhoto60 [10] dataset. Compared to competing methods, our proposed method recovers more faithful structural details (*e.g.*, the faces in (h)).

(a) LQ Patch

(b) Real-ESRGAN [7]

(c) PASD [9]

(d) DiffBIR [3]



(e) SeeSR [8]

(f) DreamClear [1]

(g) SUPIR [10]

(h) Ours

Figure 15. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers much clearer structures (*e.g.*, the restored bamboo leaves at the bottom right in (h)).

(a) LQ Patch

(b) Real-ESRGAN [7]

(c) PASD [9]

(d) DiffBIR [3]

(e) SeeSR [8]

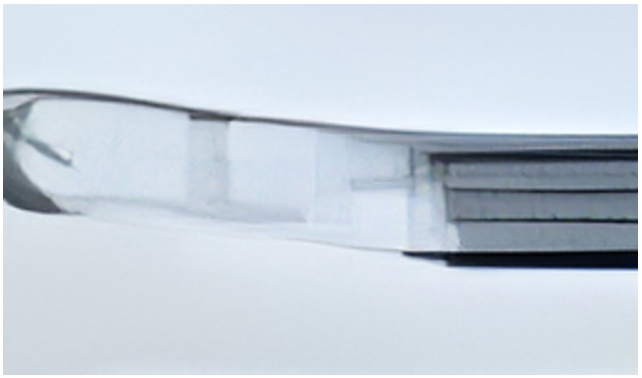
(f) DreamClear [1]

(g) SUPIR [10]

(h) Ours

Figure 16. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers realistic image with clearer structural details (*e.g.*, the restored grass, stone, and texture of the animal in (h)).

(a) LQ Patch



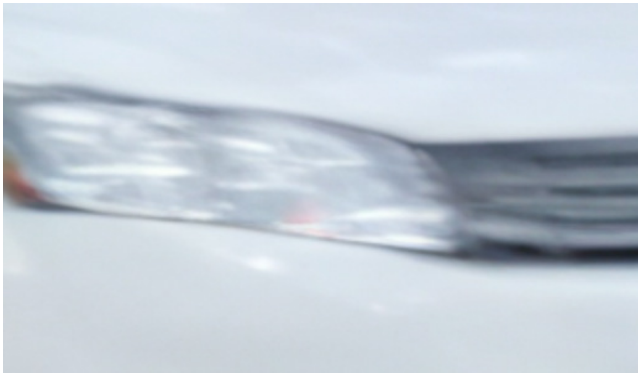
(b) Real-ESRGAN [7]

(c) PASD [9]

(d) DiffBIR [3]

(e) SeeSR [8]

(f) DreamClear [1]



(g) SUPIR [10]

(f) Ours

Figure 17. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers much clearer structure details (*e.g.*, the restored headlight and grille of a car in (f)).

(a) LQ Patch

(b) Real-ESRGAN [7]

(c) PASD [9]

(d) DiffBIR [3]

(e) SeeSR [8]

(f) DreamClear [1]

(g) SUPIR [10]

(h) Ours

Figure 18. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers much clearer structure details (*e.g.*, the restored feathers of a bird in (h)).

(a) LQ Patch

(b) Real-ESRGAN [7]

(c) PASD [9]

(d) DiffBIR [3]

(e) SeeSR [8]

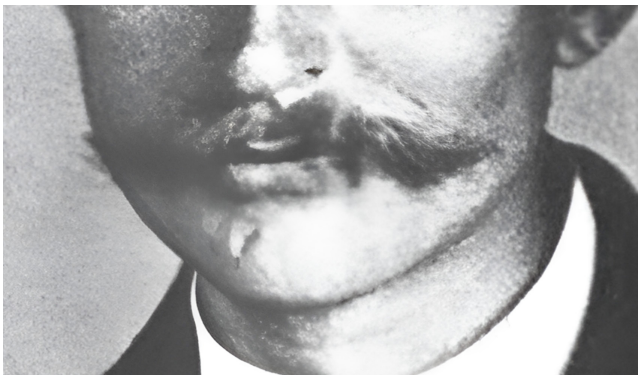
(f) DreamClear [1]

(g) SUPIR [10]

(h) Ours

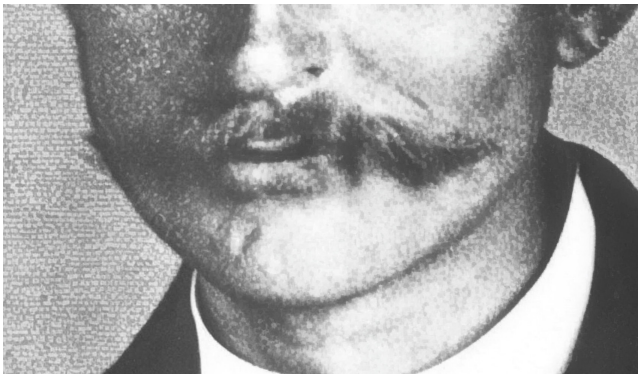
Figure 19. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers much clearer structure details (*e.g.*, the restored building and the road in (h)).

(a) LQ Patch



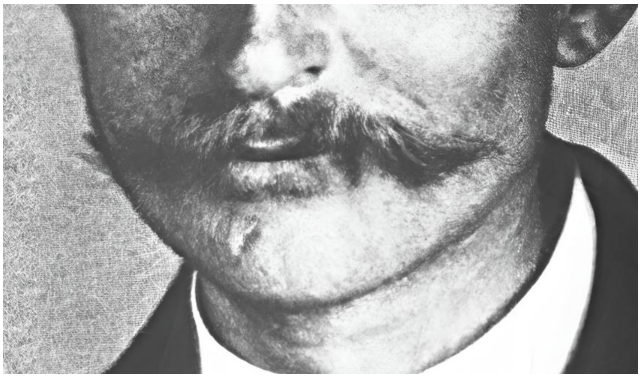
(b) Real-ESRGAN [7]

(c) PASD [9]

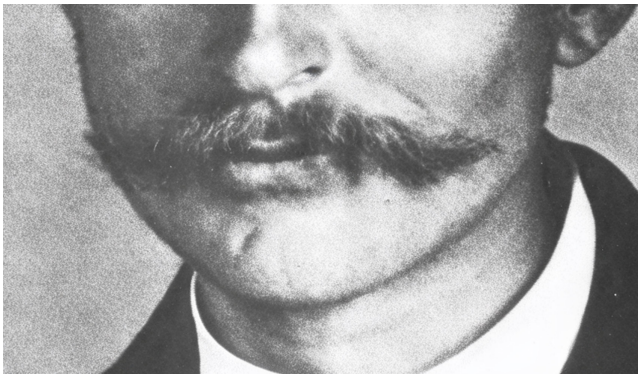


(d) DiffBIR [3]

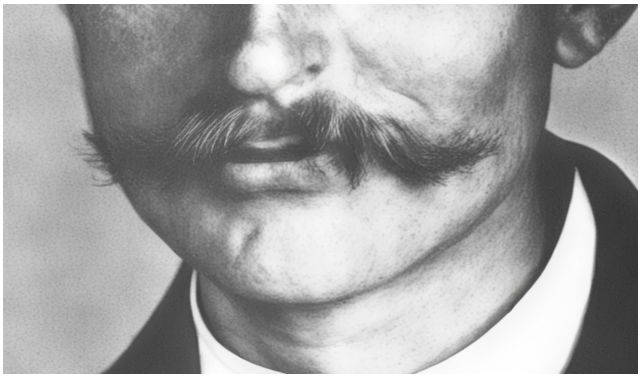
(e) SeeSR [8]



(f) DreamClear [1]



(g) SUPIR [10]



(h) Ours

Figure 20. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers more realistic details (*e.g.*, the restored beard on the face in (h)).

(a) LQ Patch

(b) Real-ESRGAN [7]

(c) PASD [9]

(d) DiffBIR [3]

(e) SeeSR [8]

(f) DreamClear [1]

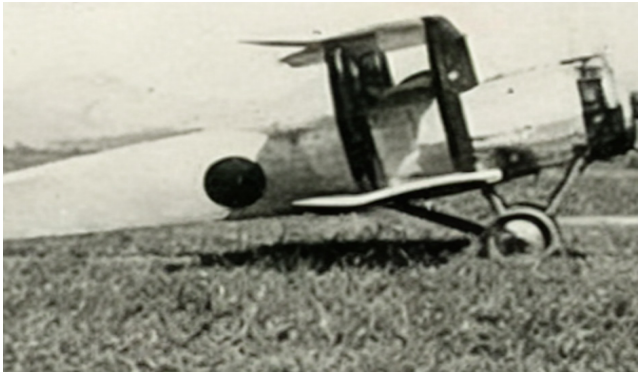
(g) SUPIR [10]

(h) Ours

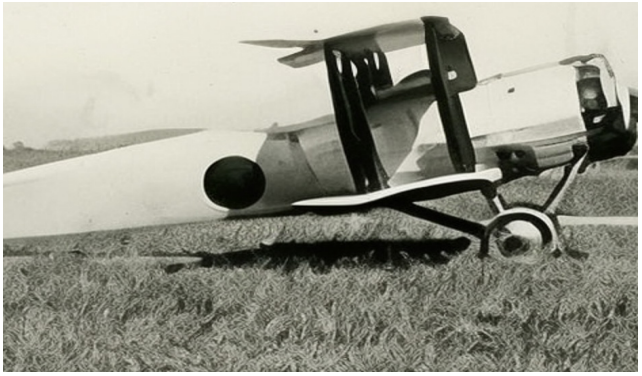
Figure 21. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers more realistic details (*e.g.*, the hair of the lion in (h)).

(a) LQ Patch

(b) Real-ESRGAN [7]



(c) PASD [9]



(d) DiffBIR [3]

(e) SeeSR [8]

(f) DreamClear [1]

(g) SUPIR [10]

(h) Ours

Figure 22. Image SR results on an example from the RealDeg dataset. Compared to competing methods, our proposed method recovers more realistic details (*e.g.*, the restored grass and plane in (h)).

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*, 2024. 3, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *CVPR*, 2021. 6

[3] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 3, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

[4] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022. 4, 5

[5] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6

[6] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024. 3, 6

[7] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *CVPR*, 2021. 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

[8] 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. 3, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

[9] 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. 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

[10] 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. 3, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

[11] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *CVPR*, 2021. 3, 6, 7, 8

[12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3