

---

# SPIDER: Boosting Blind Face Restoration via Simultaneous Prior Injection and Degradation Removal

– Supplementary Material –

---

Anonymous Author(s)

Affiliation

Address

email

1 This supplementary material provides additional details and results to support the findings presented  
2 in the main paper. It is organized into the following sections:

- 3 • **More Details on SPIDER (Section A):** We elaborate on key architectural components of our  
4 method, including the decoupled cross-attention mechanism, the training process of degradation  
5 removal module, and the degradation pipeline.
- 6 • **More Discussions on SPIDER (Section B):** We provide additional qualitative and quantitative  
7 analyses on the effectiveness of using GFPGAN for preprocessing low-quality images to obtain  
8 more accurate captions. We also examine the impact of prompt design and image captioning  
9 quality under various degradation conditions.
- 10 • **More Results on SPIDER (Section C):** This section includes extended ablation studies, addi-  
11 tional blind face restoration results under diverse degradation settings, and visualizations that  
12 demonstrate the effectiveness of both the Degradation Mapper and Remover.
- 13 • **Limitations (Section D):** We acknowledge and discuss the current limitations of our method.

## 14 A More Details on SPIDER

### 15 A.1 Details of the Decouple Cross Attention mechanisms in ControlNet.

16 ControlNet comprises several Decouple Cross Attention (DCA) blocks, each connected to corre-  
17 sponding blocks in the SD model via skip connections. These connections inject structural control  
18 into the SD model at multiple levels. This design enables structured, controllable generation by lever-  
19 aging intermediate representations from both ControlNet and the pre-trained SD pipeline. Notably,  
20 this connection strategy is inspired by the SeeSR framework [10].

### 21 A.2 Details of the Training Process of the Degradation Removal Module

22 In Stage I, we aim to encode both clean and degraded face images into a unified textural repre-  
23 sentation space. By aligning their content-consistent features, the model refines representations of  
24 degraded faces while suppressing degradation-induced noise. Notably, this process requires no ex-  
25 plicit degradation annotations, making it suitable for real-world scenarios with diverse and unknown  
26 corruptions.

27 The Degradation Removal Module (DRM) comprises two key components: the Degradation Mapper  
28 and the Degradation Remover. Both components are implemented as four-layer MLPs and trained  
29 on the FFHQ dataset [5]. For Mapper training, we use two types of inputs derived from FFHQ: the  
30 original high-quality (HQ) images and their degraded counterparts (LQ), which are generated using

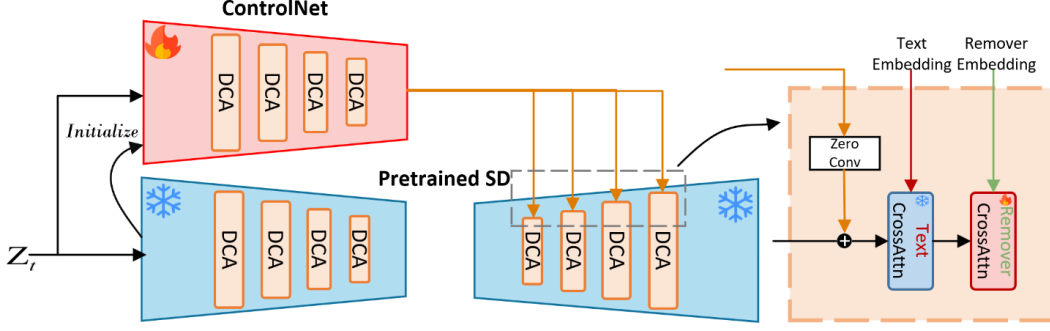


Figure 1: Overview of the integration between ControlNet and the pre-trained Stable Diffusion model.

the second-order degradation pipeline described in the previous section. The degradation parameters and sampling strategy are identical to those described in Section A.3. Notably, these HQ and LQ samples are not paired during Mapper training; instead, they are treated as unpaired data to promote generalizable representation learning under diverse and unannotated degradation conditions.

The Mapper generates a representation  $F_{\text{mapper}}$ , which is wrapped as a pseudo-text prompt in the format *"a photo of  $F_{\text{mapper}}$ "* and injected into the cross-attention layers of the pretrained Stable Diffusion model via its Key and Value branches. The Mapper is trained to reconstruct images with consistent quality (i.e., high-quality (HQ) inputs produce HQ outputs, and low-quality (LQ) inputs produce LQ outputs). Its training objective is defined as:

$$\mathcal{L}_{\text{stageI-mapper}} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0, I), t} \left[ \|\epsilon - \epsilon_\theta(z_t, t, F_{\text{mapper}})\|_2^2 \right]. \quad (1)$$

After training the Mapper, we freeze its weights and train the Degradation Remover to purify  $F_{\text{mapper}}$  by removing degradation cues, resulting in a refined embedding  $F_{\text{remover}}$  that retains content semantics while eliminating quality-related distortions. In contrast to the unpaired setting used for the Mapper, the Remover is trained using paired data, where each LQ image corresponds to its original HQ version from FFHQ. The refined representation is also wrapped as *"a photo of  $F_{\text{remover}}$ "* and injected into the same attention pathways. In this setup, the Remover receives LQ inputs and is trained to generate HQ outputs, thus enforcing a textual-level transformation guided by latent-space alignment. Its training objective follows the same diffusion loss formulation:

$$\mathcal{L}_{\text{stageI-remover}} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0, I), t} \left[ \|\epsilon - \epsilon_\theta(z_t, t, F_{\text{remover}})\|_2^2 \right]. \quad (2)$$

### A.3 Details of the Degradation Pipeline

We train all models on the FFHQ dataset [5], which contains 70,000 high-quality face images resized to  $512 \times 512$ . To generate realistic LQ training inputs, we follow the second-order degradation strategy proposed in Real-ESRGAN [9]. Each HQ image ( $y$ ) undergoes two sequential degradation processes, and the overall degradation process in one stage can be formulated as:

$$x = [(y \otimes k_\sigma) \downarrow_r + n_\delta]_{\text{JPEG}_q}, \quad (3)$$

where  $x$  is the resulting low-quality (LQ) image. The degradation includes Gaussian blur ( $k_\sigma$ ), downsampling by a factor  $r$  ( $\downarrow_r$ ), additive Gaussian noise  $n_\delta$ , and JPEG compression with quality factor  $q$  ( $\text{JPEG}_q$ ). Each degradation stage shares this structure but uses independently sampled parameters. Specifically, following DiffBIR [6], we randomly sample the degradation parameters  $\sigma$ ,  $r$ ,  $\delta$ , and  $q$  from the ranges  $[0.2, 3]$ ,  $[0.15, 1.5]$ ,  $[1, 30]$ , and  $[30, 95]$  in the first degradation stage, and  $[0.2, 1.5]$ ,  $[0.3, 1.2]$ ,  $[1, 25]$ , and  $[30, 95]$  in the second stage, respectively.

## B More Discussions on SPIDER

### B.1 Effect of Prompt Design and Image Quality on LLaVA Outputs.

The performance of LLaVA [7] is highly dependent on well-crafted prompts [8]. We use the prompt: *"Provide a detailed yet concise description of this person's face. Include their face shape, eyes, nose, mouth, eyebrows, skin texture and tone, expression, and any notable features like moles, freckles, or wrinkles."* to focus the output on detailed facial component structures. We show LLaVA outputs under three prompt types and varying degradation levels. As illustrated in Figure 2, facial description prompts yield more face-specific details, and LLaVA remains robust to various degradations.

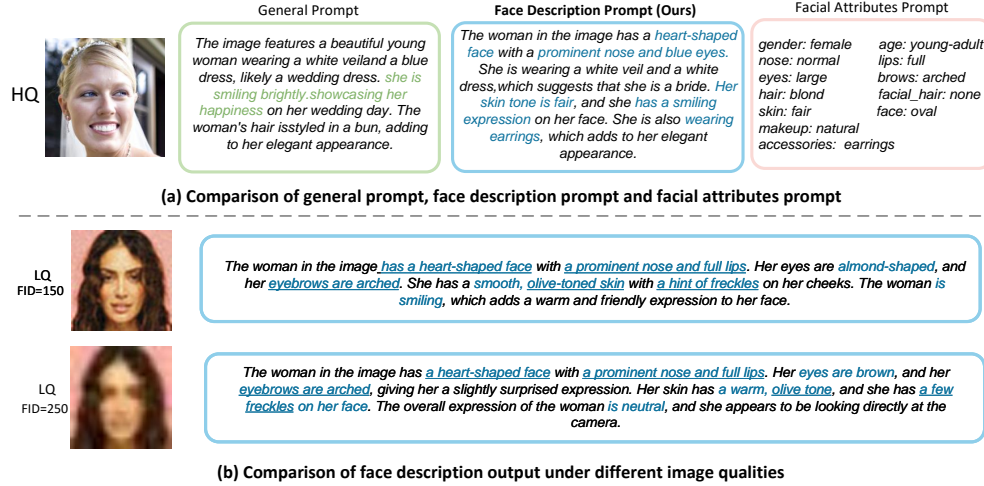


Figure 2: Comparison of different prompt types and their effects on face description across varying image qualities.

### B.2 Effect of Using GFPGAN to Preprocess Images.

Table 1 shows that using GFPGAN to preprocess low-quality images improves restoration quality. Figure 3 further demonstrates that this enhances LLaVA’s ability to generate more accurate and faithful descriptions. Similarly, FaithDiff employs BSRNet for preprocessing to improve captioning performance. Inspired by these results, we adopt GFPGAN as a preprocessing module before LLaVA to produce more precise outputs.

Table 1: A quantitative evaluation of image quality metrics on the WIDER and LFW datasets, examining the effect of applying GFPGAN as a preprocessing method for low-quality images.

Metric	WIDER		LFW	
	w/o GFPGAN	w/ GFPGAN	w/o GFPGAN	w/ GFPGAN
MUSIQ $\uparrow$	73.35	73.37	75.06	75.08
MANIQA $\uparrow$	0.5612	0.5630	0.5755	0.5784
CLIP-IQA $\uparrow$	0.7325	0.7342	0.7313	0.7320
FID $\downarrow$	35.07	34.58	39.39	39.74



Figure 3: Comparison of image restoration and description accuracy for low-quality inputs with and without GFPGAN.

## C More Results on SPIDER

### C.1 More Ablation Studies Results.

We conduct additional ablation experiments on more datasets to further validate our findings. As shown in the Table 2, on the mildly degraded LFW dataset [3], although our method exhibits slightly lower performance than other settings on a few metrics, it achieves overall the best results. On the heavily degraded WIDER dataset [11], our method significantly outperforms the other two settings, highlighting its superior robustness under complex degradation conditions.

Table 2: Ablation results showing the effectiveness of the DRM and the different module orders on WIDER and LFW datasets. The best results are marked in red, and the second best in blue.

Metric	WIDER			LFW		
	w/o DRM	DRM→Semantic	Ours	w/o DRM	DRM→Semantic	Ours
MUSIQ (↑)	70.12	71.54	73.37	75.05	75.02	75.08
MANIQA (↑)	0.5165	0.5310	0.5630	0.5983	0.5885	0.5784
CLIP-IQA (↑)	0.6971	0.7174	0.7342	0.7610	0.7556	0.7320
NIQE (↓)	5.123	5.045	5.264	5.100	5.098	5.022
FID (↓)	39.42	36.95	34.58	42.64	40.97	39.74

### C.2 More Results on Blind Face Restoration.

**More results on CelebA-Test dataset (Heavy synthetic degradations).** Figure 4 shows a visual comparison of face restoration results under severe degradation on a synthetically degraded CelebA-Test dataset. Our approach, **SPIDER (ours)**, demonstrates strong robustness to complex noise and degradation. Unlike other methods that often generate artifacts or fail to recover facial structure, SPIDER is capable of producing natural-looking faces even in extremely corrupted cases. Although identity preservation is sometimes compromised an unavoidable issue under such intense degradation the ability to restore visually plausible and realistic faces highlights the effectiveness of our model.

**More results on LFW dataset (Mild real-world degradations).** As shown in the Figure 5, for mildly degraded images, our method is able to accurately restore facial details while maintaining consistency in eye color and eye gaze.



92 **More results on Wider dataset (Heavy real-world degradations).** As shown in Figure 6, our  
 93 method reliably restores facial details under severe degradations, as evident in the first and second  
 94 rows. It preserves both structural integrity and subject identity. Additionally, as shown in the third  
 95 row, it reconstructs background subjects with high fidelity. Critical visual attributes, such as eye  
 96 color and gaze direction, are also well preserved. These results demonstrate the robustness and  
 97 semantic consistency of our approach across diverse challenging scenarios.

98 **More results on SCface dataset (Extreme surveillance degradations).** As shown in Figure 7,  
 99 Figure 8, and Figure 9, our method consistently demonstrates strong denoising capabilities across  
 100 different camera views (i.e., under various degradation conditions), while preserving facial identity  
 101 more faithfully compared to other approaches.

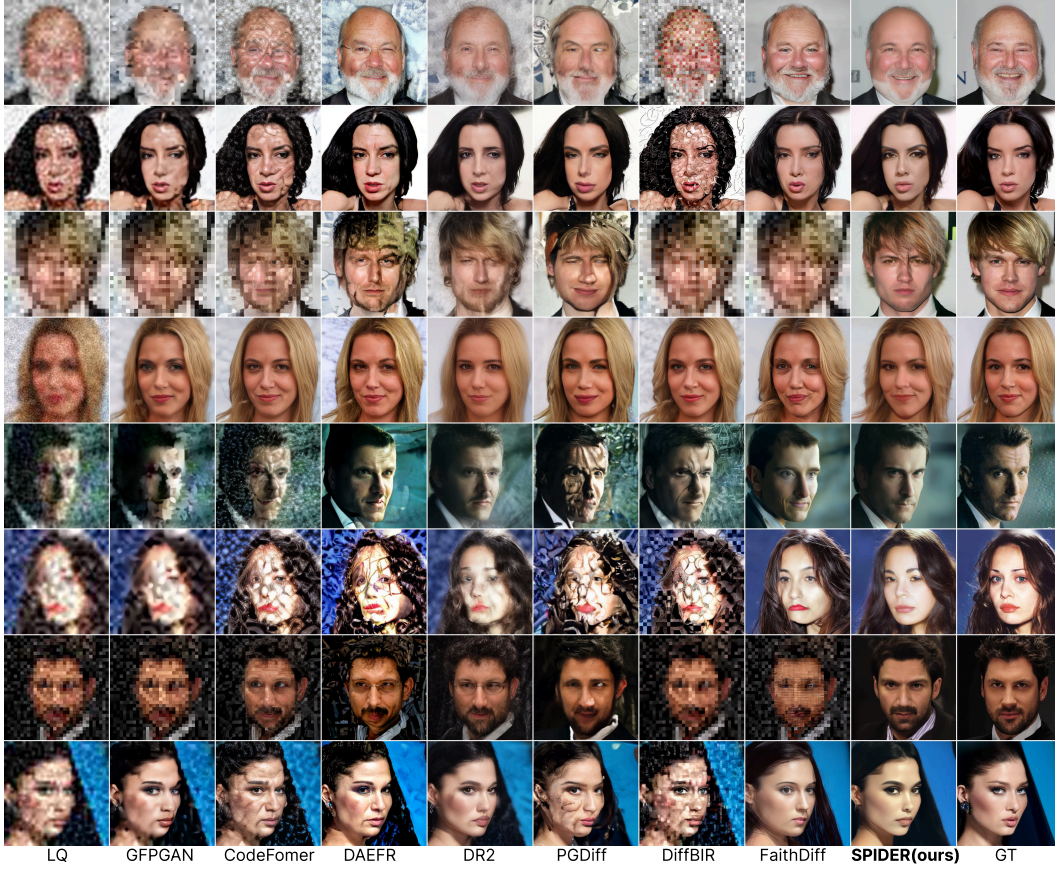


Figure 4: Visual comparison of different methods on the CelebA-Test [4] dataset.

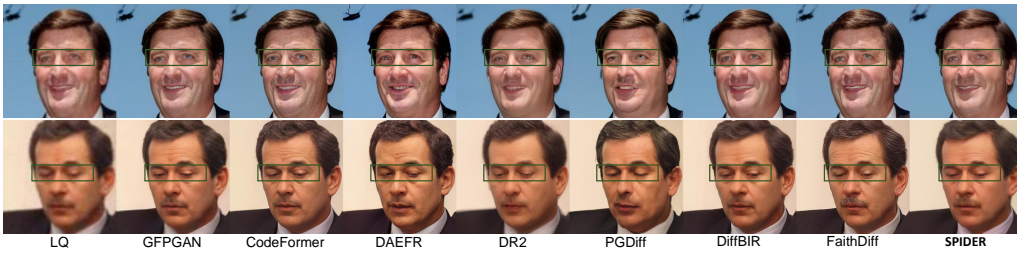


Figure 5: Visual comparison of different methods on LFW [3] dataset.



Figure 6: Visual comparison of different methods on WIDER [11] dataset.



Figure 7: Visual comparison of different methods across five camera conditions on SCface dataset [2]. (Male)





Figure 8: Visual comparison of different methods across five camera conditions on SCface dataset [2]. (Female)



Figure 9: Visual comparison of different methods on SCface dataset [2].

### 102 C.3 More Results on Degradation Removal Module.

103 Figure 10 and Figure 11 demonstrate the reconstruction ability of our Degradation Mapper trained  
 104 with unpaired data. Given either HQ or severely degraded LQ inputs, the Mapper generates seman-  
 105 tically consistent outputs, showing strong capability in encoding content and degradation patterns.  
 106 Although identity is not strictly preserved, the learned representations effectively support the subse-  
 107 quent degradation removal stage.

108 Figure 12 shows that our Degradation Remover, trained with paired LQ-HQ data, can generate high-  
 109 quality results from degraded inputs. It effectively removes various distortions and produces natural-  
 110 looking faces, demonstrating strong degradation removal capability, although identity consistency is  
 111 not strictly preserved.



Figure 10: Visual comparison of Degradation Mapper output (HQ) on the FFHQ [5] dataset.

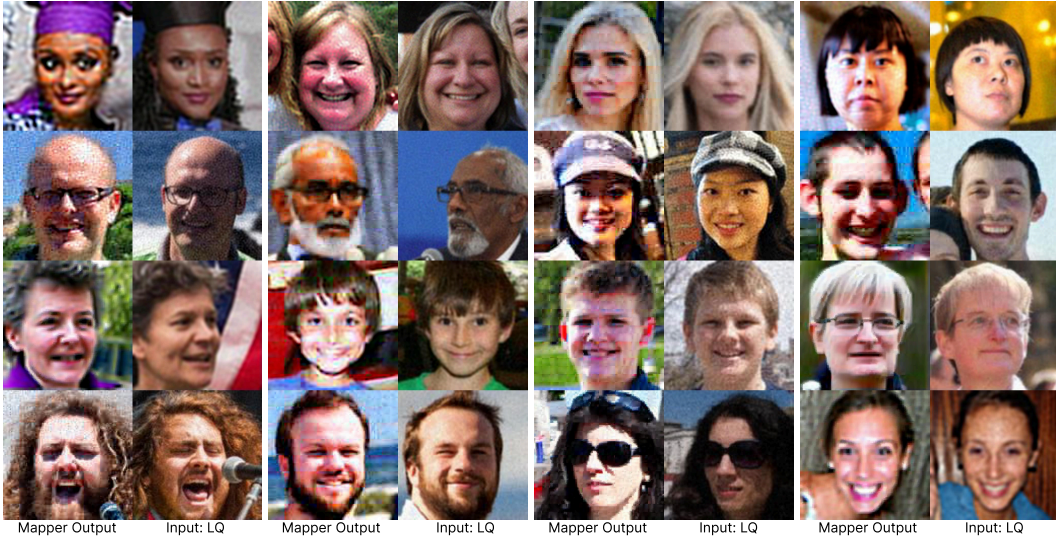


Figure 11: Visual comparison of Degradation Mapper output (LQ) on the FFHQ [5] dataset.

## D Limitations

Despite its promising performance, SPIDER still faces several limitations. First, the accuracy of image generation depends on the recognition capability of LLaVA. Under conditions of severe image degradation, LLaVA may produce inaccurate prompts, which can lead to the generation of incorrect facial attributes, as illustrated in Figure 3. This issue could potentially be mitigated by incorporating a more advanced vision-language model (VLM). In addition, since our method is built upon a pretrained Stable Diffusion model, it inherits the typical tendency of diffusion models to emphasize semantic and structural fidelity over background consistency [1], potentially resulting in slight color discrepancies in the background of the generated images.



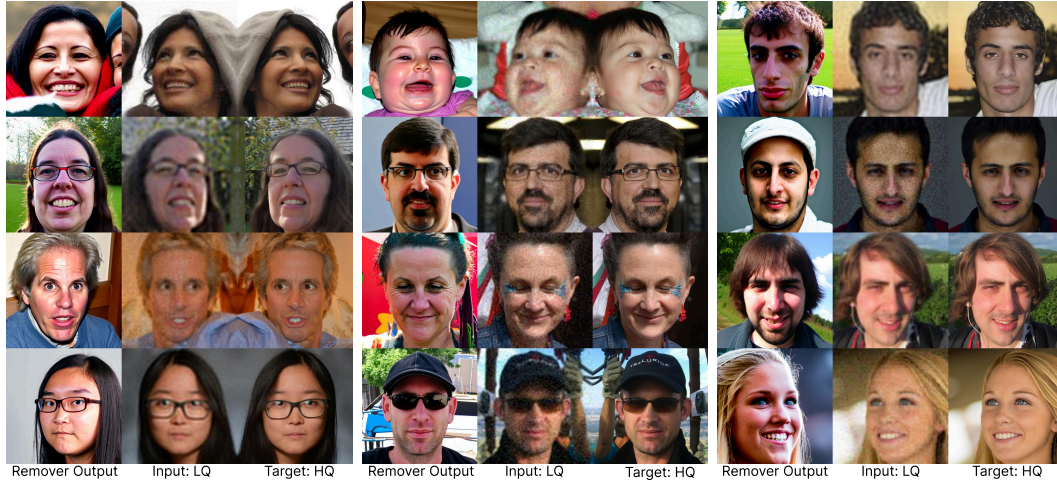


Figure 12: Visual comparison of Degradation Remover output on the FFHQ [5] dataset.

## References

- [1] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 11472–11481, 2022.
- [2] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Sface—surveillance cameras face database. *Multimedia Tools and Applications*, 51:863–879, 2011.
- [3] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces In ‘Real-Life’ Images: Detection, Alignment, and Recognition*, 2008.
- [4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations(ICLR)*, 2018.
- [5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019.
- [6] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *Proceedings of the European Conference on Computer Vision(ECCV)*, pages 430–448. Springer, 2024.
- [7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances In Neural Information Processing Systems(NeurIPS)*, 36:34892–34916, 2023.
- [8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances In Neural Information Processing Systems(NeurIPS)*, 36:34892–34916, 2023.
- [9] 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(ICCV)*, pages 1905–1914, 2021.
- [10] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, pages 25456–25467, 2024.
- [11] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 5525–5533, 2016.