LUCIDFLUX: CAPTION-FREE UNIVERSAL IMAGE RESTORATION WITH A LARGE-SCALE DIFFUSION TRANSFORMER

Song Fei[†]
The Hong Kong University of Science and Technology (Guangzhou)
sfei285@connect.hkust-gz.edu.cn

Tian Ye[†,‡]
The Hong Kong University of Science and Technology (Guangzhou)
tye610@connect.hkust-gz.edu.cn

Lei Zhu[*]
The Hong Kong University of Science and Technology
The Hong Kong University of Science and Technology (Guangzhou)
leizhu@ust.hk

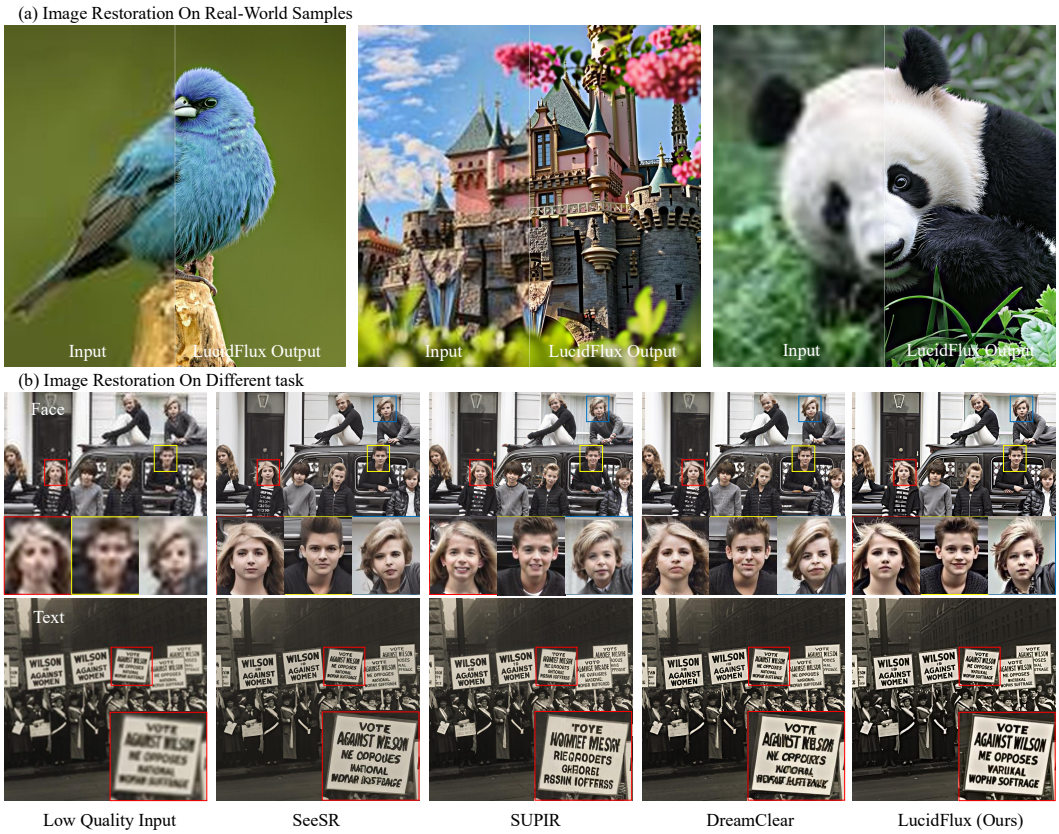

Figure 1: We present LucidFlux, a universal image restoration with a large-scale diffusion transformer that delivers photorealistic restoration of real-world LQ images, outperforming SOTA diffusion-based models in handling diverse degradations.

[†]Equal contribution [‡]Project Leader [*]Corresponding author.

ABSTRACT

Real-world image restoration is particularly challenging due to the complexity of degradations, the scarcity of paired training data, and the difficulty of balancing perceptual quality with semantic consistency. To address these issues, we develop a Flux-based diffusion framework that integrates structural priors, temporal–hierarchical modulation, and semantic alignment within a unified design. A dual-branch lightweight condition module jointly exploits information from the degraded input and a lightly restored reference, enabling robust texture recovery under severe degradations. To further enhance guidance, we introduce a timestep- and layer-adaptive modulation strategy that aligns conditional features with the diffusion process, progressively refining coarse structures into high-frequency details. Moreover, we incorporate SigLIP-based semantic priors via a connector, ensuring caption-free restoration remains semantically faithful. Extensive experiments across several real-world benchmarks demonstrate that our approach not only achieves state-of-the-art performance but also produces visually pleasing details and consistent semantics under diverse degradations.

1 METHODOLOGY

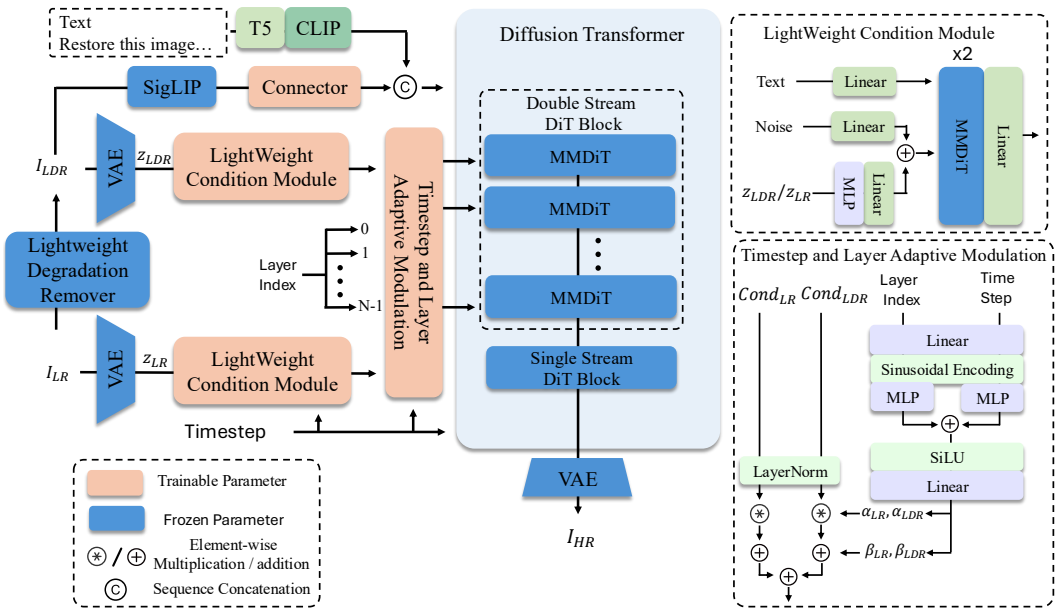

Figure 2: Overview of the proposed architecture. The model integrates dual condition streams (LR and LDR) with timestep and layer adaptive modulation modules, and incorporates SigLIP semantic priors into a Flux-based DiT backbone to jointly enhance perceptual quality and semantic consistency.

Our framework is built upon a Flux-based DiT backbone, augmented with two parallel ControlNet branches. The first branch processes the original low quality image (LR), while the second branch takes a lightly restored version of the input (LDR) generated by a lightweight restoration model. Both streams capture complementary information, which are subsequently modulated through timestep- and layer-adaptive modules to align with the DiT feature space. Moreover, we incorporate semantic priors extracted from SigLIP and enhanced with connector, which are injected into the DiT layers to facilitate semantic consistency and fine-grained texture restoration.

1.1 SCALING UP REAL-WORLD HIGH-QUALITY DATA FOR UNIVERSAL IMAGE RESTORATION

Although large-scale text-to-image (T2I) diffusion models are pretrained on hundreds of millions of image–text pairs, they are not tailored for restoration tasks. Performing post-training for image restoration requires similarly large amounts of task-relevant data. However, conventional restoration datasets are relatively small in scale; for instance, DIV2K Agustsson & Timofte (2017) contains only 800 training and 100 validation images, Flickr2K Lim et al. (2017) provides 2,650 images, and even the recently introduced LSDIR Li et al. (2023) includes about 85k images. These scales are still far below the millions of samples required for diffusion-based training.

To address this limitation, we construct a large-scale high-quality dataset using a fully automatic three-stage filtering pipeline designed to remove unsuitable samples while retaining both perceptual quality and structural richness.

Data source. Our initial dataset is collected from two sources. First, we scrape 2.3 M images from the Internet. In addition, we incorporate 557K images from the Photo-Concept-Bucket dataset bghira (2023), yielding a total of 3M candidate images. This combined pool serves as the raw data for subsequent filtering.

Blur detection. Images that are heavily blurred or contain excessive high-frequency noise provide unreliable structural cues and are thus unsuitable for training. Following LSDIR Li et al. (2023), we quantify the degree of blur using the variance of the Laplacian:

$$S_{\text{blur}}(I) = \mathrm{Var}\big(\nabla^2 I\big)\,, \tag{1}$$

where I denotes an input image. Only images with $150 \le S_{\text{blur}}(I) \le 8000$ are retained, effectively excluding both overly blurred and noisy samples.

Flat-region detection. Images dominated by textureless regions may bias the model towards producing over-smoothed outputs. To mitigate this, each image is divided into non-overlapping 240×240 patches, and the edge richness of each patch is measured using the Sobel operator:

$$S_{\text{flat}} = \mathrm{Var}\left(\sqrt{(\partial_x I)^2 + (\partial_y I)^2}\right). \tag{2}$$

Patches with $S_{\text{flat}} < 800$ are considered textureless, and images containing more than 50% such patches are discarded. This ensures that retained images exhibit sufficient edge and texture diversity, essential for high-fidelity restoration. After applying blur and flat-region filtering, 1.28 M candidate images remain.

IQA Filtering for High-quality Data. While LSDIR employs manual curation in its final stage, such human intervention is impractical for scaling to larger datasets. We apply CLIP-IQA model to further ensure perceptual quality of our training data. The remaining images are ranked by their perceptual scores s_i, and only the top 20% are retained, i.e., $\{i \mid s_i \ge \text{quantile}_{0.8}(\{s_i\})\}$, resulting in 257K high-quality images. By additionally incorporating 84K high-quality samples from LSDIR Li et al. (2023), the final curated dataset comprises 342K images.

For generating paired training data, degraded counterparts are synthesized using the Real-ESRGAN degradation pipeline Wang et al. as implemented in Ai et al. (2024), across 4 epochs, producing a total of 1.36 M image pairs. This procedure ensures both diversity and realism in the low-quality inputs, facilitating effective model training.

1.2 TWO PARALLEL LIGHTWEIGHT CONDITION MODULE FOR LOW-QUALITY IMAGE CONDITIONING

Flux.1 is originally a T2I model rather than an image-to-image (I2I) restoration network. A common adaptation is to introduce ControlNet, where two mainstream strategies exist. The first directly injects the low-quality (LR) image as a condition Chen et al. (2024), which is simple but often leaves residual artifacts under severe degradations. The second performs lightweight degradation removal before injecting features Kong et al. (2025); Yu et al. (2024), which can suppress degradations effectively but tends to oversmooth textures.

Instead of adopting either strategy in isolation, we follow the dual-branch paradigm proposed in Ai et al. (2024). One branch processes the original LR image to preserve textures, while the other

employs a Lightweight Degradation Remover (LDR) to mitigate severe degradations. This parallel design balances texture preservation and degradation suppression, providing stronger support for high-fidelity restoration:

$$I_{\text{LDR}} = \text{LDR}(I_{\text{LR}}) \,, \tag{3}$$

where I_{LR} denotes the degraded input image and I_{LDR} is the lightly restored reference image produced by the LDR.

Both images are then mapped into the shared latent space of the VAE encoder:

$$\begin{aligned} z_{\text{LDR}} &= \text{VAE}_{\text{enc}}(I_{\text{LDR}}) \,, \\ z_{\text{LR}} &= \text{VAE}_{\text{enc}}(I_{\text{LR}}) \,. \end{aligned} \tag{4}$$

Finally, the latent codes are transformed into conditioning features by the lightweight condition module (LCM):

$$\begin{aligned} \text{Cond}_{\text{LDR}} &= \text{LCM}(z_{\text{LDR}}) \,, \\ \text{Cond}_{\text{LR}} &= \text{LCM}(z_{\text{LR}}) \,. \end{aligned} \tag{5}$$

Unlike prior approaches Yu et al. (2024); Ai et al. (2024), which either trim a substantial portion of the backbone (still leaving a large parameter footprint) or duplicate it entirely, we design a lightweight condition injection module with only two MMDiT blocks. This enables effective integration of the dual-branch features while avoiding the prohibitive cost of scaling a condition module that is based on a large Flux DiT backbone. As illustrated in the top-right of Fig. 2, only the core conditioning pathway is shown for clarity, while other components such as timestep embeddings remain consistent with Flux.

1.3 Timestep and Layer-Adaptive Condition Modulation

In diffusion-based restoration, the role of the DiT backbone varies across timesteps. Early denoising stages reconstruct coarse structures, while later timesteps refine high-frequency component Park et al. (2023). Similarly, shallower layers capture low-level edges, and deeper layers process high-level semantics. Injecting identical conditional features across all layers may thus introduce redundancy or conflict.

To exploit this temporal–hierarchical structure, we apply a timestep- and layer-adaptive modulation specifically to the dual-branch MMDiT-based condition module. Let Cond_{LR} and Cond_{LDR} denote the features extracted from the low-quality input and lightly restored reference via the Lightweight Condition Module (LCM).

First, the modulation parameters α and β are predicted based on the current timestep t and layer index l:

$$\alpha^{t,l}, \beta^{t,l} = \text{Modulation}(t, l), \tag{6}$$

where $Modulation(\cdot)$ is implemented as illustrated in the lower-right of Fig. 2. Note that modulation for the LR and Reference branches is performed independently, though only one branch is visualized for clarity.

Next, the predicted modulation is applied separately to each branch:

$$\text{Cond}_{\text{LR}}^{t,l} = \alpha_{LR}^{t,l} \cdot \text{Cond}_{\text{LR}} + \beta_{LR}^{t,l}, \tag{7}$$

$$\text{Cond}_{\text{LDR}}^{t,l} = \alpha_{LDR}^{t,l} \cdot \text{Cond}_{\text{LDR}} + \beta_{LDR}^{t,l}. \tag{8}$$

Finally, the modulated features from both branches are fused to obtain the condition injected into the DiT backbone at layer l and timestep t:

$$\text{Cond}^{\text{t,l}} = \text{Cond}_{\text{LR}}^{t,l} + \text{Cond}_{\text{LDR}}^{t,l}. \tag{9}$$

By restricting modulation to the lightweight MMDiT condition module, our approach maintains efficiency while enabling coarse-to-fine, timestep- and layer-aware guidance, enhancing restoration fidelity across diverse degradations.

1.4 SIGLIP FOR CAPTION-FREE SEMANTIC ALIGNMENT

T2I diffusion models are inherently designed to synthesize images that align with textual descriptions, and thus many recent restoration methods employ captions as semantic guidance Ai et al. (2024); Yu et al. (2024). In practice, captions used for training are usually derived from the ground-truth clean images, providing an idealized semantic reference.

However, such clean captions are not available during inference, where only degraded inputs are given. Moreover, captions generated from low-quality images may introduce degradation-related artifacts into the descriptions, which can mislead the model and negatively affect restoration performance Sun et al. (2024). To alleviate this issue, several works attempt to adapt captioning models to degraded inputs. However, generating captions during inference often requires additional computation, since recent methods typically rely on large vision–language models (VLMs) such as LLaVA-13B. This not only increases inference latency but also introduces a training–inference mismatch, as the captions accessible at test time may deviate from those used during training.

To address this limitation, we propose a caption-free semantic alignment strategy. As illustrated in Fig. 2, semantic features are directly extracted from the lightly degradation removed input using SigLIP and simultaneously projected into the textual embedding space by a learnable connector:

$$z_s = \mathrm{Connector}(\mathrm{SigLIP}(I_{\mathrm{LDR}}))\,. \tag{10}$$

The projected semantics z_s are then concatenated with the text tokens c:

$$Context = \mathrm{Concat}(z_s, c), \tag{11}$$

which are fed into the DiT backbone as the multimodal conditioning context.

By grounding the restoration process on image-derived semantics, our method preserves semantic consistency while avoiding the stochastic variations often observed in T2I models, where repeated inferences under the same prompt can produce visually inconsistent results. This design enables the generation of high-fidelity outputs that are both structurally accurate and semantically aligned with the input, without requiring any external captions.

2 EXPERIMENT

2.1 IMPLEMENTATION DETAILS

All experiments are conducted on a cluster equipped with 8 NVIDIA A800 GPUs. We adopt the AdaW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01. The per-GPU batch size is set to 2, and gradient accumulation with a step size of 2 is applied, leading to an effective batch size of 32. The overall training process takes approximately 7 GPU-days. We choose SwinIR Liang et al. (2021) as our

2.2 COMPARISON WITH STATE-OF-THE-ART METHODS

We compare our method with state-of-the-art diffusion-based methods StableSR Wang et al. (2024a), SinSR Wang et al. (2024b), SeeSR Wu et al. (2024), SUPIR Yu et al. (2024), DreamClear Ai et al. (2024).

Metrics. We evaluate all methods using a set of no-reference image quality assessment metrics, including CLIP-IQA+ Wang et al. (2023), Q-AlignWu et al. (2023), MUSIQKe et al. (2021), MANIQA Yang et al. (2022), NIMA Talebi & Milanfar (2018), CLIP-IQA Wang et al. (2023), and NIQE Zhang et al. (2015). Together, these metrics provide a comprehensive assessment of restoration performance across both perceptual quality and structural fidelity.

Quantitative Comparisons. Table 1 resents quantitative results on various benchmarks. Our method achieves the highest scores on CLIP-IQA (0.7122), MUSIQ (73.01), MANIQA (0.5589), Q-Align (4.3935), NIMA (5.4836), and CLIP-IQA+ (0.7406), demonstrating consistent improvements in perceptual quality and semantic alignment. Seesr provides the second-best results across most of these metrics, confirming its competitiveness as a baseline. For NIQE, SUPIR achieves the best score of 3.6591, while our method ranks second with 3.6742. Overall, the results highlight

Table 1: Quantitative comparison across different IQA metrics on RealSR Wu et al. (2024) and RealLQ250 Ai et al. (2024).

Benchmark	Metric	ResShift	StableSR	SinSR	SeeSR	DreamClear	SUPIR	LucidFlux(Ours)
RealSR	CLIP-IQA+ ↑	0.5005	0.4408	0.5416	0.6731	0.5331	0.5640	**0.7074**
	Q-Align ↑	3.1045	2.5087	3.3615	3.6073	3.0044	3.4682	**3.7555**
	MUSIQ ↑	49.50	39.98	57.95	67.57	49.48	55.68	**70.20**
	MANIQA ↑	0.2976	0.2356	0.3753	0.5087	0.3092	0.3426	**0.5437**
	NIMA ↑	4.7026	4.3639	4.8282	4.8957	4.4948	4.6401	**5.1072**
	CLIP-IQA ↑	0.5283	0.3521	0.6601	**0.6993**	0.5390	0.4857	0.6783
	NIQE ↓	9.0674	6.8733	6.4682	5.4594	5.2873	5.2819	**4.2893**
RealLQ250	CLIP-IQA+ ↑	0.5529	0.5804	0.6054	0.7034	0.6810	0.6532	**0.7406**
	Q-Align ↑	3.6318	3.5586	3.7451	4.1423	4.0640	4.1347	**4.3935**
	MUSIQ ↑	59.50	57.25	65.45	70.38	67.08	65.81	**73.01**
	MANIQA ↑	0.3397	0.2937	0.4230	0.4895	0.4400	0.3826	**0.5589**
	NIMA ↑	5.0624	5.0538	5.2397	5.3146	5.2200	5.0806	**5.4836**
	CLIP-IQA ↑	0.6129	0.5160	**0.7166**	0.7063	0.6950	0.5767	0.7122
	NIQE ↓	6.6326	4.6236	5.4425	4.4383	3.8700	**3.6591**	3.6742

Figure 3: Qualitative comparison on RealLQ250. While baseline methods either leave noticeable artifacts or yield oversmoothed textures, our approach restores sharper details. Best viewed with zoom.

that our approach delivers superior performance on perceptual quality metrics while maintaining competitive scores on distortion-oriented measures.

Qualitative Comparisons. Figure 3 presents visual comparisons on representative samples from RealLQ250. Seesr and DreamClear reduce some degradations but tend to leave residual artifacts or produce oversmoothed outputs with limited texture recovery. SUPIR generates cleaner results yet often loses fine details, leading to overly smooth surfaces. In contrast, our method achieves clearer edges, richer textures, and better semantic consistency with the degraded inputs, especially

in challenging regions such as hair, text, and high-frequency patterns. These qualitative observations align with the quantitative results in Table 1, further confirming the effectiveness of our approach.

3 CONCLUSION

We presented a Flux-based universal image restoration framework that integrates dual lightweight condition branches, timestep and layer-adaptive modulation, and semantic prior from Siglip for caption-free semantic alignment. Complemented by a fully automatic data filtering pipeline and the integration of LSDIR and Real-ESRGAN degradations, our approach leverages 338K curated high-quality samples to enable data-efficient training. Experiments on RealSR and RealLQ250 demonstrate that our method consistently restores sharper textures and achieves superior perceptual quality compared to existing baselines.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017.

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. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024.

bghira. Photo concept bucket, 2023. URL https://huggingface.co/datasets/bghira/photo-concept-bucket. Accessed: 2025-09-05.

Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. In *European conference on computer vision*. Springer, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers, 2025. URL https://arxiv.org/abs/2504.17825.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

Haoze Sun, Wenbo Li, Jiayue Liu, Kaiwen Zhou, Yongqiang Chen, Yong Guo, Yanwei Li, Renjing Pei, Long Peng, and Yujiu Yang. Text boosts generalization: A plug-and-play captioner for real-world image restoration, 2024. URL https://openreview.net/forum?id=RjwWClPZtV.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

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

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

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*, pp. 25456–25467, 2024.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191–1200, 2022.

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, 2024.

Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. doi: 10.1109/TIP.2015.2426416.

A Appendix

You may include other additional sections here.