

INSTANTIR: BLIND IMAGE RESTORATION WITH INSTANT GENERATIVE REFERENCE

Jen-Yuan Huang^{1, 2} Haofan Wang² Qixun Wang² Xu Bai^{2, 3}
Hao Ai² Peng Xing² Jen-Tse Huang⁴
¹Peking University ²InstantX Team ³Xiaohongshu Inc
⁴The Chinese University of Hong Kong

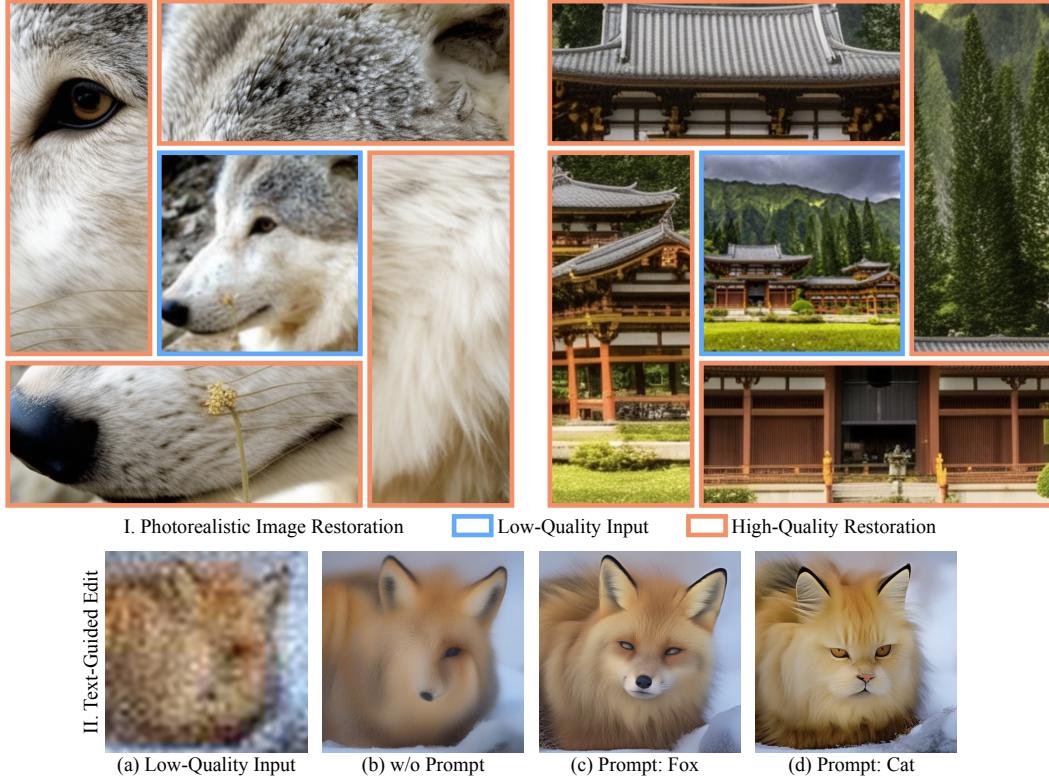


Figure 1: I. INSTANTIR presents exceptional capability in reproducing photorealistic details. II. INSTANTIR provides an active interface for natural language guidance, helps handling large degradation and features creative restoration with semantic editing.

ABSTRACT

Handling test-time unknown degradation is the major challenge in Blind Image Restoration (BIR), necessitating high model generalization. An effective strategy is to incorporate prior knowledge, either from human input or generative model. In this paper, we introduce Instant-reference Image Restoration (INSTANTIR), a novel diffusion-based BIR method which dynamically adjusts generation condition during inference. We first extract a compact representation of the input via a pre-trained vision encoder. At each generation step, this representation is used to decode current diffusion latent and instantiate it in the generative prior. The degraded image is then encoded with this reference, providing robust generation condition. We observe the variance of generative references fluctuate with degradation intensity, which we further leverage as an indicator for developing a sampling algorithm adaptive to input quality. Extensive experiments demonstrate INSTANTIR achieves state-of-the-art performance and offering outstanding visual quality.

Through modulating generative references with textual description, INSTANTIR can restore extreme degradation and additionally feature creative restoration.

1 INTRODUCTION

Image restoration seeks to recover High-Quality (HQ) visual details from Low-Quality (LQ) images. This technology has a wide range of important applications. It can enhance social media contents to improve user experience (Chao et al., 2023). It also functions at the heart in industries like autonomous driving (Patil et al., 2023) and robotics (Porav et al., 2019) by improving adaptability in diverse environments, as well as assists object detector in adverse conditions (Sun et al., 2022).

Image restoration remains a long-standing challenge extending beyond its practical application. The information loss during degradation makes a single LQ image corresponds to multiple plausible restorations. This ill-posed problem is further exacerbated in Blind Image Restoration (BIR), where models are tested under unknown degradation. A common strategy is to leverage prior knowledge. Reference-IR models use other HQ images to modulate LQ features, requiring additional inputs with similar contents but richer visual details (Lu et al., 2021). Generative approaches, on the other hand, directly learn the HQ image distribution. The input is first encoded into a hidden variables z , which servers as the generation condition to sample HQ image from the learned distribution $p(y|z)$. Although generative methods achieve single-image restoration, they are prone to hallucinations that produce artifacts in restoration (Yang et al., 2020). This happens when the encoder fails to retrieve accurate hidden variable due to the input distribution shift in degradation. Existing methods improve robustness by training on more diverse synthetic degradation data or introduce discrete feature codebook. We argue that these are only shot-term solutions. Alternative methods are pending to be explored to better address unknown inputs in BIR.

In this paper, we present INSTANTIR, a dynamic restoration pipeline that iteratively refines generation condition using a pre-trained Diffusion Probabilistic Model (DPM). INSTANTIR employs two complementary way for processing input LQ image. First, a pre-trained vision encoder extracts compact representation from degraded content. The encoder’s high compression rate enhances the robustness in the extracted representation, while retaining only high-level semantics and structural information. Next, we introduce the *Previewer* module, a distilled DPM capable of one-step generation. At each generation step, the previewer decodes current diffusion latent using the compact representation, providing a restoration preview resembles original input in high-level features. This preview serves as an instant generative reference to guide the *Aggregator* in encoding identity and other fine-grained missing from the compact representation. We observe in experiments that the previewer tends to decode aggressively when the input is clear, resulting in high variance in restoration previews. We take this as a reliable indicator of input image quality, and develop an adaptive sampling algorithm that amplifies the fine-grained encoding with relatively high quality inputs. Additionally, we find the previewer is controllable through text prompts, which produces diverse generative references and enables semantic editing with restoration. Our contributions are as follows:

1. We explore a novel BIR method that iteratively aligns with the generative prior to address unknown degradation;
2. We introduce a novel architecture based on pre-trained DPM, which dynamically adjusts the generation condition by previewing intermediate outputs;
3. We develop sampling algorithms tailored for our pipeline, enabling both adaptive and controllable restoration to text prompts;
4. We perform extensive evaluations to validate the effectiveness of the proposed methods.

2 RELATED WORK

2.1 DIFFUSION MODEL

DPM is a class of generative model that generate data by iteratively denoising from Gaussian noise (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b). Typically, a neural network with a UNet architecture (Ronneberger et al., 2015) is trained to predict the noise added at

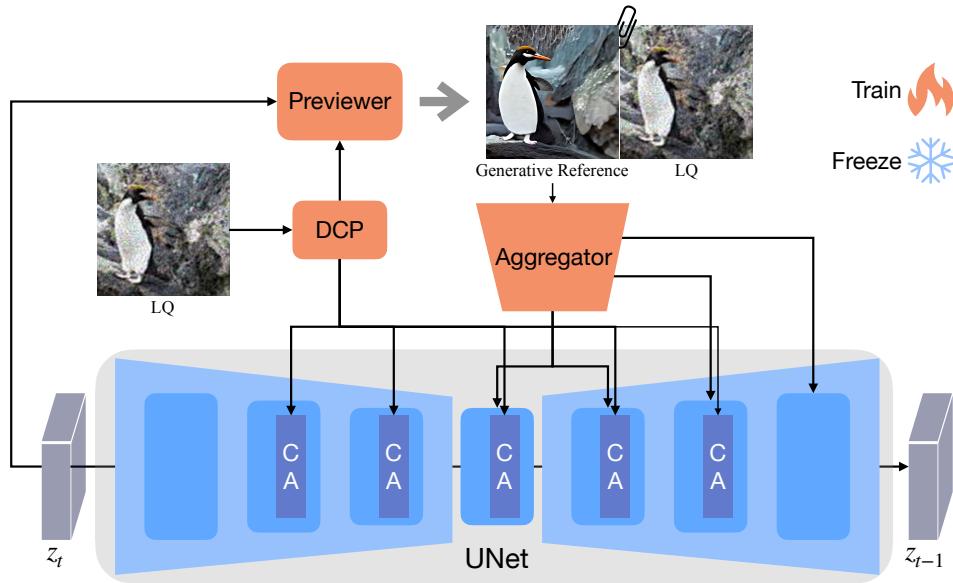


Figure 2: The overall pipeline of INSTANTIR. INSTANTIR adopts a novel previewing mechanism to actively align with generative prior. This is achieved by three key modules: 1) DCP for compact LQ image representation encoding; 2) Previewer for decoding it into generative prior; and 3) Aggregator for integrating the generative reference and LQ input into sampling conditions.

each inference step. DPM offers superior mode coverage compared to Variational Autoencoders (VAE) (Kingma & Welling, 2013) and outperform GAN-based models (Goodfellow et al., 2020) in generation quality without the need of adversarial training (Dhariwal & Nichol, 2021). These advantages establish DPM as the leading approach in vision generative models. By incorporating additional inputs, DPMs can learn diverse conditional distributions (Nichol & Dhariwal, 2021), with the most widely used application being text-to-image (T2I) generation (Rombach et al., 2022; Saharia et al., 2022a; Ramesh et al., 2022). Leveraging the flexibility of text inputs and the vast amount of text-image training data (Schuhmann et al., 2022), these models are capable of generating images with exceptional visual quality and remarkable diversity, forming the foundation for many subsequent excellent work in vision generative models (Wang et al., 2024c;a).

2.2 BLIND IMAGE RESTORATION

The task setting makes BIR particular valuable in real-world applications. The major challenge in BIR is the input distribution gap between training and testing data. Previous work have explored multiple ways to address this issue. Feature quantification is widely used in generative-based methods (Esser et al., 2021; Van Den Oord et al., 2017; Zhou et al., 2022). They align the encoded LQ image features to a learnable feature codebook, ensuring the input to generator is unaffected by domain shifts. However, this hard alignment constraints the generation diversity and quality by the capacity of the discrete codebook. Previous work have also explored the application of powerful DPM in BIR. Some approaches design specialized architectures and train DPMs from scratch (Saharia et al., 2022b; Sahak et al., 2023; Li et al., 2022), while the others apply additional modules on pre-trained T2I model (Wang et al., 2024b; Yu et al., 2024; Sun et al., 2024a), leveraging their large-scale prior. In many practical scenarios, HQ images with similar contents, such as those from photo albums or video frames, are available. This has spurred interest in restoring images using reference-based methods (Cao et al., 2022; Jiang et al., 2021; Lu et al., 2021; Xia et al., 2022; Yang et al., 2020; Zhang et al., 2019). They adopt regression models to learn how to transfer high-quality features to LQ images, enhancing details restoration.

3 METHODOLOGY

The distribution gap between training and testing data exacerbates the ill-posed nature of BIR, causing hallucinations in generation-based IR models and producing artifacts. We attribute this to the error in encoding LQ image, and propose a generative restoration pipeline that refines the LQ encodings with generative references. This is achieved by exploiting the reverse process of DPM. Specifically, we first encode the LQ image into a compact representation via pre-trained vision encoder, capturing global structure and semantics to initiate diffusion generation. Conditioned on this embedding, our Previewer module generates a restoration preview at each diffusion time-step. The preview resembles to the input image with more plausible details, and they are further fused in the Aggregator module to preserve fidelity. Finally, the adjusted LQ encoding is used to control the pre-trained DPM for a fine-grained diffusion step.

3.1 PRELIMINARIES

DPM involves two stochastic processes named forward and reverse process (Ho et al., 2020). In the forward process, *i.i.d.* Gaussian noise is progressively added to the image \mathbf{x} . The marginal distribution of diffusion latent \mathbf{x}_t follows $\mathcal{N}(\alpha_t \mathbf{x}, \beta_t \mathbf{I})$, where α_t and β_t are hyperparameters defining the forward process. \mathbf{x}_t converges to pure noise as t increases, and the reverse process generates images by inverting the forward process. Generally, we train a neural-network to predict the noise added at each time-step by minimizing the diffusion loss:

$$\mathcal{L}_{diff} = \mathbb{E} [\|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon\|^2], \quad (1)$$

where ϵ_θ denotes the noise-prediction network. At each step in the reverse process, we can retrieve a denoising sample with the predicted noise and re-parameterization (Karras et al., 2022):

$$\hat{\mathbf{x}} = \frac{\mathbf{x}_t - \beta_t \epsilon_\theta(\mathbf{x}_t, t)}{\alpha_t}. \quad (2)$$

In the open-sourced T2I model Stable Diffusion (SD) (Rombach et al., 2022), the noise-prediction network ϵ_θ is additionally conditioned on a text input that describes the target image. Moreover, SD employs a VAE to move the input \mathbf{x}_t into latent space \mathbf{z}_t , compressing inputs by a factor of 48 and significantly reduces the memory usage to enable image generation up to 512^2 resolution.

3.2 ARCHITECTURE

The restoration pipeline of INSTANTIR consists of three key modules: Degradation Content Perceptor (DCP) for compact LQ image encoding, Instant Restoration Previewer for generating references on-the-fly during the reverse process, and Latent Aggregator for integrating restoration references.

Degradation Content Perceptor We employ the pre-trained DINO (Oquab et al., 2023) for providing compact LQ image representation. Compared to CLIP (Radford et al., 2021), a common choice in image editing (Ye et al., 2023), DINO’s self-supervised training with data augmentation improves robustness of the encoded features. The extracted LQ representation is modulated by a learnable Resampler (Han et al., 2024) and projected as context to the cross-attention layers of diffusion UNet. For the l -th cross-attention block, we introduce an additional cross-attention operation:

$$\mathbf{f}_{out}^l = \mathbf{f}_{in}^l + \text{CrossAttn}(\mathbf{f}_{in}^l, \mathbf{c}_{txt}) + w^l \cdot \text{CrossAttn}(\mathbf{f}_{in}^l, \Phi(\mathbf{c}_{lq}, t)), \quad (3)$$

where Φ denotes the DCP module and \mathbf{c}_{lq} is the LQ context matrix. We retain the text cross-attention here as it is a crucial part of the pre-trained T2I model that synthesizes high-level semantics. Jointly training DCP with textual transformation allows it to focus on low-level information absent in the other modality. We introduce a hyper-parameter w^l to regulate their behaviors. Note that the DCP also takes time-step t as input to establish temporal dependency in the output. Specifically, we use adaptive layer-normalization to modulate the context matrix from the DCP according to time-step t :

$$\Phi(\mathbf{x}, t) = \mathcal{T}_{scale} \odot \text{LayerNorm}(\mathbf{c}_{lq}) + \mathcal{T}_{shift}, \quad (4)$$

where, $\mathcal{T}_{scale}, \mathcal{T}_{shift}$ are calculated from the time-step. We train the DCP module on a frozen diffusion model using the standard diffusion loss in Eq. 1.

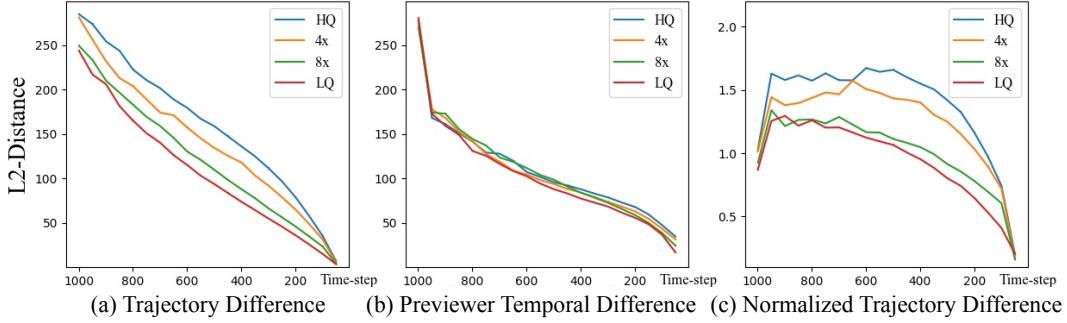


Figure 3: The evolution of the Previewer outputs during generation. (a) L2-distances between previews and denoising means; (b) temporal differences of the Previewer trajectory, measured by L2-distances between adjacent points; (c) relative distances between previews and denoising means.

Instant Restoration Previewer The compact representation encoded by the DCP, while robust against degradation, loses high-level information. We introduce Previewer, a diffusion model generates from current diffusion latent instead of noise, to decode generative references from the DCP encoding. Decoding at each diffusion time-step requires $(T(T+1)/2)$ network forward passes with the vanilla T2I model. To streamline this process, we fine-tune the Previewer using consistency distillation (Luo et al., 2023) to make it a one-step generator. For diffusion latent \mathbf{z}_s at time-step s , we first obtain the Previewer output conditioned solely on \mathbf{c}_{lq} . Then, we perform a diffusion step using the pre-trained model from \mathbf{z}_s , conditioned on both \mathbf{c}_{lq} and \mathbf{c}_{txt} , to reach \mathbf{z}_t . \mathbf{z}_t is regarded as the ground-truth diffusion latent at time-step t in the sampling trajectory. Finally, we get the preview of \mathbf{z}_t , again conditioned solely on \mathbf{c}_{lq} . The consistency distillation loss is then calculated by:

$$\mathcal{L}_{dist} = \|\Psi(\mathbf{z}_s, s, \Phi(\mathbf{c}_{lq}, s)) - \text{StopGrad}(\Psi(\mathbf{z}_t, t, \Phi(\mathbf{c}_{lq}, t)))\|^2, \quad (5)$$

where Ψ denotes the previewer model. Additionally, Eq. 5 trains the previewer to follow the sampling trajectory without \mathbf{c}_{txt} , removing its dependency on text conditions which are typically unavailable in BIR tasks. The consistency constraint (Song et al., 2023) of enforcing consistent outputs across time-step enabling the Previewer to decode generative references on-the-fly.

Latent Aggregator The primary challenge in the BIR task is the input distribution shift. Previous work address this by aligning LQ features with reference HQ images or a learned feature codebook. The former takes extra inputs, while the latter is limited to a specific domain by the codebook capacity. In contrast, we generate reference features directly from diffusion prior. Since the compact embedding \mathbf{c}_{lq} retains only high-level information, it is insufficient for the Previewer to reconstruct HQ images at larger time-steps, as shown in Fig. 6. Relying solely on reference preview incurs error accumulation, so the Aggregator anchors preview to the original input to prevent divergence in the reverse process. The input LQ image is encoded into SD’s latent space and spatially concatenated with the preview. This expanded input remains compatible to the diffusion UNet, allowing the Aggregator to be initialized as a trainable copy of UNet compression path following (Zhang et al., 2023). We remove text cross-attention layers to make the Aggregator lightweight and independent of textual conditions like the Previewer. The preview and LQ hidden features are fused in the spatial-attention layers, which are further integrated via Spatial Feature Transform (SFT) (Wang et al., 2018). For hidden feature \mathbf{H}^l at the l -th layer in the Aggregator, we first split it spatially into \mathbf{h}_p^l and \mathbf{h}_o^l , corresponding to the hidden features of preview and LQ latent, and integrate them with SFT:

$$\mathbf{h}_{res}^l = (1 + \boldsymbol{\alpha}^l) \odot \mathbf{h}_p^l + \boldsymbol{\beta}^l; \mathbf{h}_p^l, \mathbf{h}_o^l = \text{Split}(\mathbf{H}^l), \quad (6)$$

where $\boldsymbol{\alpha}^l, \boldsymbol{\beta}^l = \mathcal{M}_\theta^l(\mathbf{h}_o^l)$ are two affine transformation parameters calculated from the feature map of LQ latent at this level. We extract multi-level features $\{\mathbf{h}_{res}^l\}_{l=1}^L$ from Aggregator using Eq. 6, and inject them into the corresponding part of U-Net expansion path through residual connections.

3.3 ADAPTIVE RESTORATION

INSTANTIR processes LQ image through two complementary ways: 1) extracting compact representation using the DCP, which is robust to degradation but loses fine-grained information; 2) encoding

via the lossless SD-VAE and integrating with restoration preview, which is prone to errors in the SD-VAE. Under severe degradation, INSTANTIR may produce samples deviate from the target HQ image. In such cases, restoration previews exhibit small variation, suggesting the DCP struggles to provide guidance according to the input. We further analyze the trajectory of restoration previews during the reverse process, compare it with the denoising predictions from Eq. 2. We assess them on four degradation levels: HQ image, 4x downsampling, 8x downsampling and synthetic multi-degradation, representing decreasing input quality. Fig. 3 (a) illustrates the L2-distance between these two trajectories, which increases monotonically as input quality improves. A pronounced disparity between preview and ordinary denoising prediction represents the Previewer is confident with the guidance, suggesting the input LQ image is informative. Based on this observation, we use the relative difference between two predictions as an indicator of input quality:

$$\delta = \frac{\|\Psi(\mathbf{z}_t, t, \Phi(\mathbf{c}_{lq}, t)) - \hat{\mathbf{z}}_t\|^2}{\|\Psi(\mathbf{z}_t, t, \Phi(\mathbf{c}_{lq}, t)) - \Psi(\mathbf{z}_{t+1}, t+1, \Phi(\mathbf{c}_{lq}, t+1))\|^2}, \quad (7)$$

where $\hat{\mathbf{z}}_t$ is given by Eq. 2. From Fig. 3 (b) we can see the Previewer is unstable at the beginning. The consistency training in Eq. 5 drives it to decode aggressively, causing large prediction variance during early reverse process where the input diffusion latent is too noisy. Normalizing the L2-distance between trajectories with Previewer’s temporal difference effectively mitigates the temporal correlation as illustrated in Fig. 3 (c). A larger δ indicates higher input quality, and the conditional signals from the Aggregator should be amplified to preserve fine-grained information from the original input. On the other hand, DPM is known to first generate low-frequency features such as global structure, and add high-frequency details in the later stage of the reverse process. A decreasing δ encourages INSTANTIR to produce diverse by exploiting the LQ image representation. We provide pseudo-code of the proposed adaptive restoration (AdaRes) algorithm in Alg. 1.

Surprisingly, although only the DCP module is explicitly trained on text-image data, INSTANTIR demonstrates notable creativity following textual descriptions. By employing a text-guided Previewer, we can generate diverse restoration variations with compound semantics from both modalities. However, these variation samples can conflict with the original input, making them ineligible as generative references. Inspired by previous work in image editing, we disable the Aggregator at later stage generation and let INSTANTIR renders semantic details according to LQ representation and text prompt. This ensures the low-frequency features are succeeded from the Aggregator, meanwhile prevents the high-frequency semantics and noise from entering the final results.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

INSTANTIR is built on SDXL (Podell et al., 2023) accompanied by a two-stage training strategy. In Stage-I, we train the DCP module on a frozen SDXL, followed by consistency distillation of the Previewer (see Sec. 3.2). The Previewer is trained by applying Low-Rank Adaptation (LoRA) (Hu et al., 2021) on the base SDXL model for efficiency. By toggling the Previewer LoRA, we can seamlessly switch between the Previewer and SDXL, reducing memory footprint. After obtaining the DCP and Previewer LoRA, we proceed to Stage-II Aggregator training. The two-stage training ensures the Aggregator receives high-quality previews since the beginning of its training course.

We adopt SDXL’s data preprocessing and conduct training on 1024^2 resolution. In both two stages we use the AdamW (Loshchilov, 2017) optimizer with a learning rate of 1×10^{-4} . In Stage-I, we train the DCP module using a batch size of 256 over 200K steps, and distill the Previewer for another 30K steps with the same batch size. We train the Aggregator with a batch size of 96 over 200K steps in Stage-II. The entire training process spans approximately 9 days on 8 Nvidia H800 GPUs.

Table 1: Quantitative comparisons on both synthetic validation data and public real-world dataset. We highlight the best results in **bold** and the second best with underline.

Dataset	Model	PSNR	SSIM	LPIPS	CLIPQA	MANIQA	MUSIQ
Synthetic	BSRGAN	20.21	0.5214	0.7793	0.2072	0.2076	17.53
	Real-ESRGAN	19.92	<u>0.5317</u>	0.7554	0.2102	0.2331	17.39
	StableSR	<u>20.42</u>	0.5388	<u>0.3751</u>	0.4672	0.2602	52.33
	CoSeR	19.92	0.5114	0.3353	0.6651	<u>0.4152</u>	<u>67.51</u>
	SUPIR	20.46	0.4990	0.4090	0.4875	0.3081	56.43
	INSTANTIR (ours)	18.54	0.5126	0.3986	<u>0.5497</u>	0.4379	68.59
Real-world	BSRGAN	26.38	0.7651	0.4120	0.3151	0.2147	28.58
	Real-ESRGAN	27.29	0.7894	0.4173	0.2532	0.2398	25.66
	StableSR	26.40	<u>0.7721</u>	0.2597	0.4501	0.2947	48.79
	CoSeR	25.59	0.7402	<u>0.2788</u>	0.5809	<u>0.3941</u>	<u>60.51</u>
	SUPIR	<u>26.41</u>	0.7358	0.3639	0.3869	0.2721	42.72
	INSTANTIR (ours)	21.75	0.6766	0.3686	<u>0.5401</u>	0.4819	65.32

(a) Scenario 1: 512^2 image restoration. The outputs of SUPIR and INSTANTIR are downsampled to 512^2 .

Dataset	Model	PSNR	SSIM	LPIPS	CLIPQA	MANIQA	MUSIQ
Synthetic	BSRGAN	21.32	<u>0.5267</u>	0.5611	0.4289	0.3299	37.97
	Real-ESRGAN	20.45	0.5202	0.5660	0.4566	0.3627	37.92
	StableSR	<u>21.01</u>	0.5490	0.3921	0.4526	0.2492	48.94
	CoSeR	20.50	0.5215	0.3488	0.6461	0.3939	<u>64.84</u>
	SUPIR	20.57	0.4569	0.4196	<u>0.6286</u>	<u>0.3962</u>	61.00
	INSTANTIR (Ours)	18.80	0.5076	<u>0.3903</u>	0.6111	0.4303	66.09
Real-world	BSRGAN	28.60	<u>0.8141</u>	0.3690	0.4720	0.2258	18.26
	Real-ESRGAN	<u>28.13</u>	0.8209	0.3647	0.4435	0.3229	35.31
	StableSR	27.79	0.8043	0.2514	0.4634	0.2901	46.54
	CoSeR	27.04	0.7683	<u>0.2882</u>	0.5847	<u>0.4068</u>	<u>58.39</u>
	SUPIR	26.10	0.5825	0.5429	0.4822	0.3232	44.95
	INSTANTIR (Ours)	21.89	0.6879	0.3601	<u>0.5647</u>	0.4389	62.58

(b) Scenario 2: 1024^2 image restoration. We crop 512^2 patches as inputs to 512 -models and evaluate the quantitative metrics on the cropped area only.

To enable Classifier-free Guidance (CFG) (Ho & Salimans, 2022) sampling, we apply LQ image dropout with a probability of 15% in both stages training. In all test experiments, we employ 30 steps DDIM sampling (Song et al., 2020a) with a CFG scale of 7.0.

4.2 EXPERIMENTAL CONFIGURATION

Training Data We synthesis LQ-HQ image pairs using Real-ESRGAN (Wang et al., 2021) with the default setting. As mentioned in Sec. 3.2, we conduct Stage-I training on the JourneyDB dataset (Sun et al., 2024b), a generated dataset with descriptive captions. While JourneyDB images are of extreme quality, they lack the textures in real-world images. Hence for Stage-II training, we incorporate publicly available texture-rich datasets to enhance model’s ability to produce realistic visual details. Specifically, we use DIV2K (Agustsson & Timofte, 2017), LSDIR (Li et al., 2023), Flickr2K (Timofte et al., 2017) and FFHQ (Karras et al., 2019).

Test Setting For a comprehensive evaluation, we test INSTANTIR on a synthetic dataset and public benchmarks following previous work. We synthesize 2,000 multi-degradation samples from DIV2K and LSDIR validation sets using Real-ESRGAN pipeline, filtering out images smaller than 1024^2 to ensure ground-truth quality. We include a small portion of JourneyDB validation data to enhance benchmark diversity. We conduct evaluations on RealSR (Cai et al., 2019) and DRealSR (Wei et al., 2020) to assess model performance on real-world LQ images. We report full-reference metrics PSNR, SSIM, LPIPS (Zhang et al., 2018), if ground-truth targets are available, and non-reference



Figure 4: Qualitative comparisons on real-world LQ images. Restorations from INSTANTIR are rich in details with global semantic consistency. Better viewed zoom in.

metrics MANIQA (Yang et al., 2022), CLIPQA (Wang et al., 2023), MUSIQ (Ke et al., 2021) to quantitatively compare INSTANTIR with other models.

4.3 COMPARING TO EXISTING METHODS

We compare INSTANTIR with state-of-the-art models, including StableSR (Wang et al., 2024b), CoSeR (Sun et al., 2024a), SUPIR (Yu et al., 2024), BSRGAN (Zhang et al., 2021) and Real-ESRGAN (Wang et al., 2021). Since some of them are limited to 512^2 resolution, we consider two test scenarios for a fair comparison: 1) models are tested on 512^2 images with outputs of 1024^2 -models scaled accordingly; 2) following SUPIR, the models are tested on 1024^2 images by cropping 512^2 patch as inputs to 512 -models, metrics are evaluated on the cropped area only.

Quantitative Comparison The results are summarized in Tab. 1. INSTANTIR continuously achieves the highest MUSIQ and MANIQA scores across all test settings, outperforming the second best by large margins up to **22%** in MANIQA and **8%** in MUSIQ. Notably in scenario 1, despite halving the input data, INSTANTIR still performs comparably to SOTA models. While CoSeR achieves the best CLIPQA scores closely followed by INSTANTIR, restorations from 1024 -models SUPIR and INSTANTIR are rich in details as shown in Fig. 4. We also observe the misalignment of PSNR and SSIM scores with visual quality as reported in the literature (Yu et al., 2024; Wang et al., 2024b). We include these metrics here for reference purpose.

Qualitative Comparison We provide some restoration samples on real-world LQ images in Fig. 4. Through leveraging the previewing mechanism, INSTANTIR actively aligns with generative prior, reducing hallucinations and producing sharp yet realistic details. In the second row of Fig. 4, while SUPIR’s result contains rich textures, the absence of global semantic guidance causes the diver’s body and mask to blend together. In contrast, the cognitive encoder in CoSeR helps it identify statues in the second example. CoSeR employs a feature codebook to handle unknown degradations, which limits the generation of complex textures on the statues. Notably in the first row of Fig. 4,

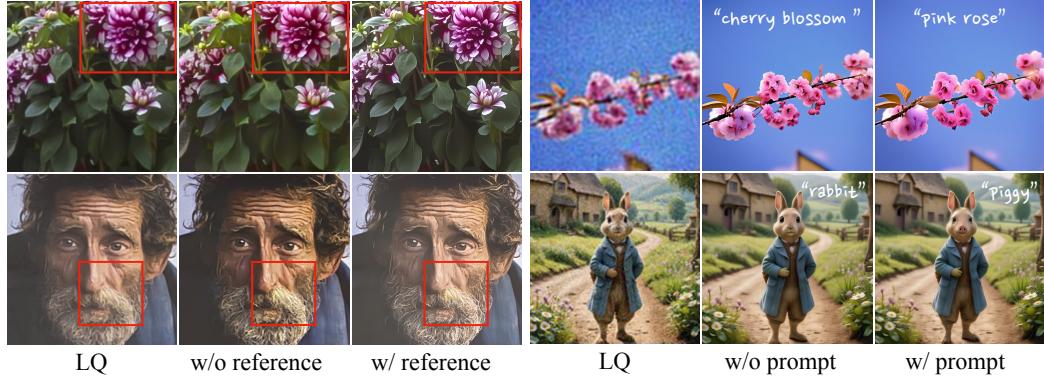


Figure 5: Visual examples of the previewing mechanism in INSTANTIR. Better viewed zoom in.

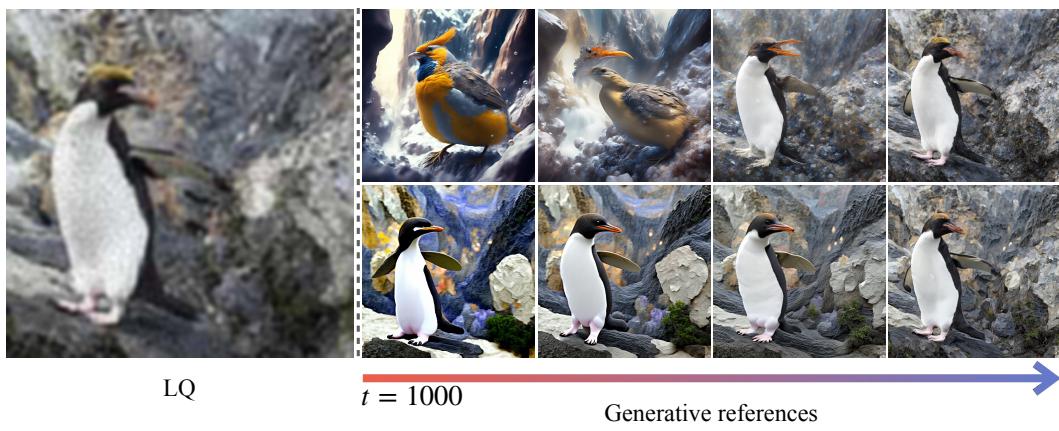


Figure 6: Visual examples of generative references. The first row on the right is generated by DCP trained on image-only data, while the second row is produced by DCP used in INSTANTIR, which is trained on text-image pairs.

INSTANTIR is the only one that successfully recovers all four faces without distortion, suggesting its superior ability in capturing semantic and reproduce realistic details from diverse degradations.

4.4 RESTORATION WITH PREVIEWING

In-domain Reference for Detail Enhancement Reference-based BIR models improve detail restoration by transferring high-quality textures from HQ references. INSTANTIR achieves this by querying the T2I model, eliminating additional inputs. In Fig. 5a, we disable the Previewer to see the effect of generative references. Here INSTANTIR infers solely with LQ images, which is beneficial to fidelity preservation but bad for visual quality. This is also reflected in Tab. 2b where all quantitative metrics deteriorate except PSNR and SSIM. Moreover, INSTANTIR equipped with Alg. 1 further improves the non-reference metrics, suggesting its flexibility in to different conditions.

Out-domain Reference for Creative Restoration Fig. 5b shows more creative restoration samples. Owing to the efficiency of our Aggregator in integrating reference latents, INSTANTIR is able to perform high-level semantic editing during restoration, altering specific attributes of the subject and leaving other visual details like global structure and layout intact. We empirically find INSTANTIR offer better prompt-following ability under heavy degradation.

Table 2: Ablation studies. The best results are highlighted in **bold**.

	PSNR	SSIM	LPIPS	CLIPQA	MANIQA	MUSIQ
Baseline	21.40	0.6775	0.3173	0.5433	0.4024	66.35
+Distillation	24.24	0.6963	0.4306	0.2453	0.2145	38.33
+Noisy Previews	23.07	0.7312	0.3830	0.3767	0.2924	49.23

(a) Ablation study of the consistency distillation in Previewer and adding fresh noise to restoration previews.

References	AdaRes	PSNR	SSIM	LPIPS	CLIPQA	MANIQA	MUSIQ
✗	✗	22.24	0.7539	0.3672	0.2721	0.2128	42.64
✓	✗	21.13	0.6728	0.3173	0.5445	0.3747	64.86
✓	✓	21.06	0.6708	0.3189	0.5456	0.3766	64.94

(b) Ablation study of the generative references and AdaRes sampling.

4.5 ABLATION STUDY

DCP Training on Text Domain We compare training DCP module with and without textual condition. Due to limited computational resources, we did not proceed to train the subsequent Previewer and Aggregator for the DCP trained on image-only data. For comparison, we provide some visual examples of their generative references across diffusion time-steps in Fig. 6. As shown in the first row of Fig. 6, the generative references from the image-only DCP differ significantly from the input LQ image at early stage, retaining only coarse semantic like “a bird standing on a rocky surface.” In contrast, DCP trained with text descriptions preserves most of the low-level information, including global hue, structure, layout, and even the subject’s category (penguin) and its pose.

Previewer Consistency Distillation We validate the necessity of consistency constraints in Previewer. We experiment with using predictions from Eq. 2 as reference inputs to the Aggregator. The second row in Tab. 2b shows a significant drop in the non-reference metrics. In fact, the prediction in Eq. 2 is close to the distribution mean at each time-step (Karras et al., 2022). Previewer with consistency distillation can directly sample from the data distribution, providing more informative generative references.

Fresh Noise to Restoration Previews We additionally train an Aggregator that injects fresh noise to reference latents according to diffusion time-step. The noisy preview latent follows the same distribution as current diffusion latent, making the overall pipeline resemble a ControlNet model (Zhang et al., 2023). As shown in the third row of Tab. 2b, INSTANTIR significantly outperforms ControlNet with LQ image as conditional inputs. This highlights the effectiveness of the previewing mechanism in INSTANTIR for adjusting generation conditions during inference.

5 CONCLUSION

In this paper, we explore a novel method to address unknown degradations in BIR tasks. Through exploiting the generation process of DPM, we propose to actively align with the generative prior to reduce the errors in encoding LQ image. Our pipeline is implemented based on pre-trained SDXL model, referred to as INSTANTIR. Extensive experiments demonstrate the exceptional restoration capability of INSTANTIR, delivering SOTA performance in quantitative metrics and visual quality. However, we observe some disparity in reference metrics such as PSNR and SSIM compared to SOTA models, which might because of the excessive generative prior diminishes fidelity. Future work will explore approach to improve the interaction between generative prior and conditions, as well as ways to refine the previewer to produce more reliable references.

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*, pages 1–10. IEEE, 2017.

nition workshops, pp. 126–135, 2017.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3086–3095, 2019.

Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, pp. 325–342. Springer, 2022.

Jiahao Chao, Zhou Zhou, Hongfan Gao, Jiali Gong, Zhengfeng Yang, Zhenbing Zeng, and Lydia Dehbi. Equivalent transformation and dual stream network construction for mobile image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14102–14111, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts. *arXiv preprint arXiv:2406.09162*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2103–2112, 2021.

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*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

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.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

- Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. 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.
- I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2021.
- Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.
- Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Prashant W Patil, Sunil Gupta, Santu Rana, Svetha Venkatesh, and Subrahmanyam Murala. Multi-weather image restoration via domain translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21696–21705, 2023.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Horia Porav, Tom Bruls, and Paul Newman. I can see clearly now: Image restoration via de-raining. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7087–7093. IEEE, 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- 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*, pp. 10684–10695, 2022.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.
- Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv preprint arXiv:2302.07864*, 2023.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022a.
- Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022b.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.
- Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.
- Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25868–25878, 2024a.
- Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aoju Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024b.
- Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *Advances in Neural Information Processing Systems*, 35:4461–4474, 2022.
- Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 114–125, 2017.
- Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024a.
- Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 2 in 37, pp. 2555–2563, 2023.
- Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pp. 1–21, 2024b.
- Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024c.
- Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 606–615, 2018.
- 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*, pp. 1905–1914, 2021.
- 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*, pp. 101–117. Springer, 2020.
- Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 3 in 36, pp. 2768–2776, 2022.

- Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5791–5800, 2020.
- 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.
- Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.
- Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4791–4800, 2021.
- 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*, pp. 3836–3847, 2023.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.
- Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7982–7991, 2019.
- Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.