This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Acquire and then Adapt: Squeezing out Text-to-Image Model for Image Restoration

Junyuan Deng[1] Xinyi Wu[1] Yongxing Yang[1] Congchao Zhu[1] Song Wang[1,2] Zhenyao Wu[1*]
[1]Honor Device Co., Ltd [2]Shenzhen University of Advanced Technology
{dengjunyuan1, wuxinyi, yangyongxing1, zhucongchao, wangsong5, wuzhenyao}@honor.com

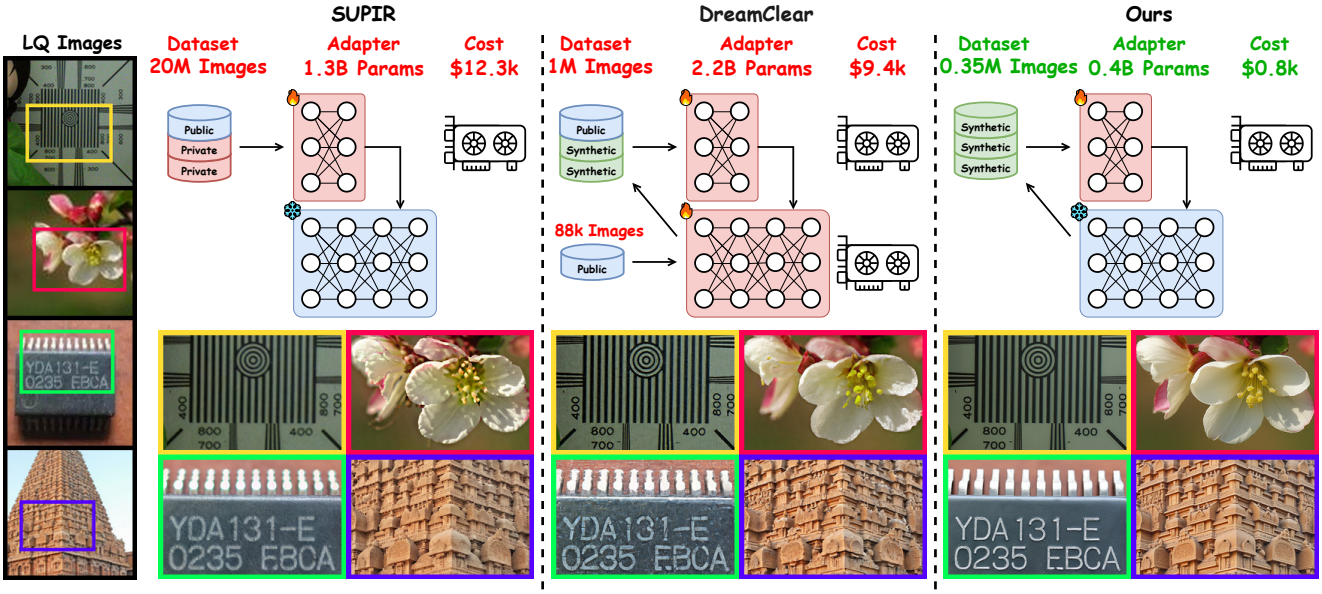

Figure 1. Comparison of SUPIR [91], DreamClear [3], and our proposed method. Our training dataset is constructed entirely from synthetic images. Trained with such data, our method achieves the most realistic restoration results with the lowest training cost.

Abstract

Recently, pre-trained text-to-image (T2I) models have been extensively adopted for real-world image restoration because of their powerful generative prior. However, controlling these large models for image restoration usually requires a large number of high-quality images and immense computational resources for training, which is costly and not privacy-friendly. In this paper, we find that the well-trained large T2I model (i.e., Flux) is able to produce a variety of high-quality images aligned with real-world distributions, offering an unlimited supply of training samples to mitigate the above issue. Specifically, we proposed a training data construction pipeline for image restoration, namely FluxGen, which includes unconditional image generation, image selection, and degraded image simulation. A novel light-weighted adapter (FluxIR) with squeeze-and-excitation layers is also carefully designed to control the large Diffusion Transformer (DiT)-based T2I model so that reasonable details can be restored. Experiments demonstrate that our proposed method enables the Flux model to adapt effectively to real-world image restoration tasks, achieving superior scores and visual quality on both synthetic and real-world degradation datasets - at only about 8.5% of the training cost compared to current approaches.

1. Introduction

In real-world scenarios, images often suffer from diverse and unpredictable degradations during capture, storage and transmission, giving rise to a wide range of image restoration (IR) tasks, *e.g.*, deblurring [12, 34, 54, 59, 67, 80, 98], denoising [7, 77, 96, 97], super-resolution [10, 17, 29, 37], and *etc.*, for restoring the original high-quality images based on the type of degradation. These methods, however, strug-

*Corresponding author.

gle to maintain generalizability under complex real-world degradation due to the limited model capacity and constrained training data. To alleviate this problem, many researchers have developed methods [44, 73, 83, 91] based on pre-trained text-to-image (T2I) models [18, 50, 51, 60]. Trained on vast collections of real image-text pairs, these T2I models encapsulate extensive prior knowledge, enabling them to enrich low-quality images of any type with realistic details. However, an immense dataset and extensive training period are required to prevent alterations of image content and the generation of inaccurate details.

Early generative-based IR methods [44, 47, 52, 53, 65, 66, 73, 83, 103] are typically built upon the Stable-Diffusion model, combining several publicly available datasets such as DIV2K [2], Flickr2K [42], LSDIR [39] and DIV8K [24] to create a diverse training set. As generative models have scaled up, the following work [91] increased the scale of training data by collecting more high-quality, high-resolution images. Nevertheless, acquiring large amounts of data remains challenging: manual data collection incurs high costs, while web scraping poses potential copyright and privacy risks. Furthermore, the resulting datasets remain limited in scope. The recent work [3] approaches these challenges by incorporating one million generated high-quality images. However, the utilization of the sophisticated multimodal large language model (*i.e.*, Gemini-1.5-Pro [68]) and the necessity for finetuning the T2I model result in a less efficient data curation process. In this paper, we propose **FluxGen**, a streamlined and highly efficient data generation pipeline that operates without the involvement of LLMs. Specifically, we find Flux [36] can generate highly realistic and diverse images directly from random noise. We also incorporate no-reference image quality assessments (IQA) methods [31, 72, 88] for image filtering, along with advanced degradation techniques [75, 99] for constructing training image pairs.

Typically, for generative-based IR methods [44, 73, 83], large-scale trainable adapters are needed to control T2I models external control signals. These adapters are usually replications of the U-Net [61] encoder from the T2I model or the full model, resulting in substantial training time and GPU resource demands. Moreover, with the continued growth in the parameter size of T2I models, recent efforts have concentrated on scaling up adapters to more effectively control these increasingly powerful models. For example, SUPIR [91] employs a 1.3B adapter for 3.5B SDXL [51], while DreamClear [3] introduces 2.2B adapter for 0.6B PixArt-α [9]. In this work, we carefully design a light-weighted ControlNet-like adapter **FluxIR** with only 0.4B trainable parameters to adapt the powerful 12B pre-trained Flux [36] model to the IR task. The adapter consists of an MM-DIT [18] block and multiple squeeze-and-excitation (SE) layers, where the former extracts control signal features and the latter modulates the denoising process efficiently using these features. As illustrated in Fig. 1, our proposed method reduces training costs[1] by approximately 93.5% and 91.5% compared to SUPIR [91] and DreamClear [3], respectively, while also achieving the best image restoration quality. The main contributions of this paper are summarized as follows:

- We are the first to confirm that the Flux model exhibits the remarkable capacity to generate highly lifelike images for the curation of IR datasets. Our proposed FluxGen can automatically generate an unlimited number of training samples in a simple and effective manner.
- We design FluxIR, a light-weighted adapter equipped with a few squeeze-and-excitation layers, that can effectively manipulate the super-large T2I model like Flux to restore reasonable details for degraded images within 14 GPU-days.
- Extensive experiments demonstrate that our method achieves optimal performance in handling real-world degraded images and producing the most realistic and satisfactory results.

2. Related Work

Image Restoration. Image restoration aims to reconstruct clear, detailed images from degraded ones under various types of degradation. Early approaches often assumed specific degradation processes [14, 16, 26, 29, 37, 62, 101] or incorporated degradation prediction directly into the model [20, 49, 71, 74, 79, 95]. However, these methods struggle to generalize to real-world scenarios. Recent studies have advanced image restoration by building complex real-world degradation processes [75, 99]. These approaches simulate real-world degradation through varied random combinations of factors like noise, blur, JPEG compression, *etc*. These simulations have driven notable improvements in restoration performance, boosting image quality substantially [6, 40, 41, 75, 99]. Building on the successes of text-to-image diffusion models [18, 36, 51, 60], several studies [8, 11, 13, 19, 23, 30, 38, 70, 73, 82, 83, 86, 89, 91] have leveraged the generative prior from pre-trained T2I models to restore images with significantly enhanced quality and high-frequency details. In this work, we take a further step by leveraging the generative prior from the advanced pre-trained T2I diffusion model, Flux, and introduce a novel MM-DiT-based adapter to recover missing details and enhance the aesthetic quality of the input images.

Large Text-to-Image Diffusion Model. Diffusion Model [15, 27, 63, 64] have achieved superior performance on image synthesis. Stable Diffusion (SD) [60] is one of the most famous models, which compresses images into

[1]Training costs are calculated by multiplying the GPU rental price by the total GPU hours used.

the latent space using a high-fidelity high-compression ratio VAE [35]. Later, Diffusion transformer (DiT) [50] proposed replacing the traditional U-Net framework [61] with a transformer, aiming to improve scalability when trained on large datasets. As a result, SD3 [18], which adopts the multimodal diffusion transformer-based (MM-DiT) framework, has demonstrated outstanding performance in both text-to-image generation and visual quality. More recently, Flux [36] extended the MM-DiT [18] architecture and used a 16-channel VAE to scale the T2I diffusion model to 12B parameters. By integrating the T5 [56] text encoder, Flux achieved top performance in text-image alignment, visual quality, and aesthetic quality. In this paper, we leverage the generative prior of Flux to build a high-quality IR training dataset and restore the details of degraded images.

Data Distillation. Scaling up datasets has proven essential for improving the performance of large-scale models across various fields, including natural language processing (NLP) [1, 69], text-to-image [18, 36, 51, 60], image restoration [91], and *etc*. However, collecting such large-scale real-world data is prohibitively expensive. As a result, researchers are increasingly turning to generated data, distilling datasets from generative models to make dataset expansion more efficient and cost-effective. For instance, BOOT [22] and DKDM [85] utilize synthetic data generated from pretrained models to optimize the diffusion models. GSDD [94] employs GAN inversion techniques [84] to learn a GAN distribution and generate images for the image super-resolution task. Recently, 800K training samples generated by DeepSeek-R1 [25] have been used to enhance the reasoning capabilities of small dense models like Qwen [87] and Llama[21]. Closest to our work, DreamClear [3] fine-tunes a text-to-image diffusion model to generate high-quality data with designed prompts. However, both [94] and [3] still require external data for either tuning or inversion, which is time-consuming and raises privacy concerns. To address this, we propose a novel data generation pipeline that distills a high-quality IR training dataset directly from pre-trained text-to-image models, without the need for external data and additional training.

3. Method

In Sec. 3.1, we introduce FluxGen, a pipeline for generating high-quality, realistic training images using Flux. Then, in Sec. 3.2, we present FluxIR, a ControlNet-like adapter that enables precise control over Flux for image restoration. Finally, Sec. 3.3 details efficient training strategies, including a novel timestep sampler and pixel-space loss functions.

3.1. FluxGen

Image generation. Both the quality and volume of training images are critical for neural network performance in image restoration tasks. Existing works [58, 91, 102] choose to collect millions of images to build their training data. However, constructing such large datasets introduces four significant challenges: 1) The data collection process is labor-intensive, requiring extensive human effort for preprocessing; 2) The use of this data raises privacy and copyright concerns; 3) Commercially purchased datasets are prohibitively expensive, limiting access for many research institutions; 4) Handling large volumes of high-resolution images is difficult due to high bandwidth and storage requirements.

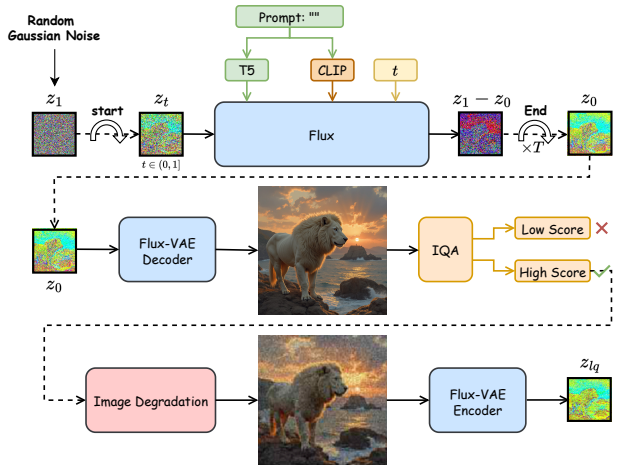

Figure 2. An overview of our FluxGen pipeline. First, an empty prompt and random Gaussian noise z_1 are input into Flux, generating an image latent z_0 over T steps. A VAE decoder then maps z_0 to its corresponding image x_0. High-quality images are curated by IQA-based selection, followed by image degradation to construct the final paired dataset.

Recently, Flux has effectively leveraged the scalable MM-DiT [18] alongside a 16-channel VAE to achieve leading performance in generating high-quality, high-resolution images. These images display realistic high-frequency texture details that far surpass those produced by other T2I models. In our proposed method, we utilize Flux for the first time to generate a large volume of high-quality training images at a low cost. Unlike [58, 91, 102], acquiring high-quality and realistic images from Flux is labor-free, copyright-compliant, cost-effective, and easy to implement.

As shown in Fig. 2, we input an empty prompt into Flux and manipulate the initial random noise to generate millions of images using the capabilities of the pre-trained Flux model. Here we take advantage of the random text prompt-dropping design in diffusion-based text-to-image (T2I) models [18, 57, 60] which allows T2I models to produce a diverse range of realistic images from their generative prior without the designed prompt. Notably, we can also leverage large language models (LLMs) like GPT-4 [1] and LLaMA [69] to craft tailored text prompts for FluxGen. The corresponding generated images can be used to train image restoration models focused on specific objects (*e.g.*,

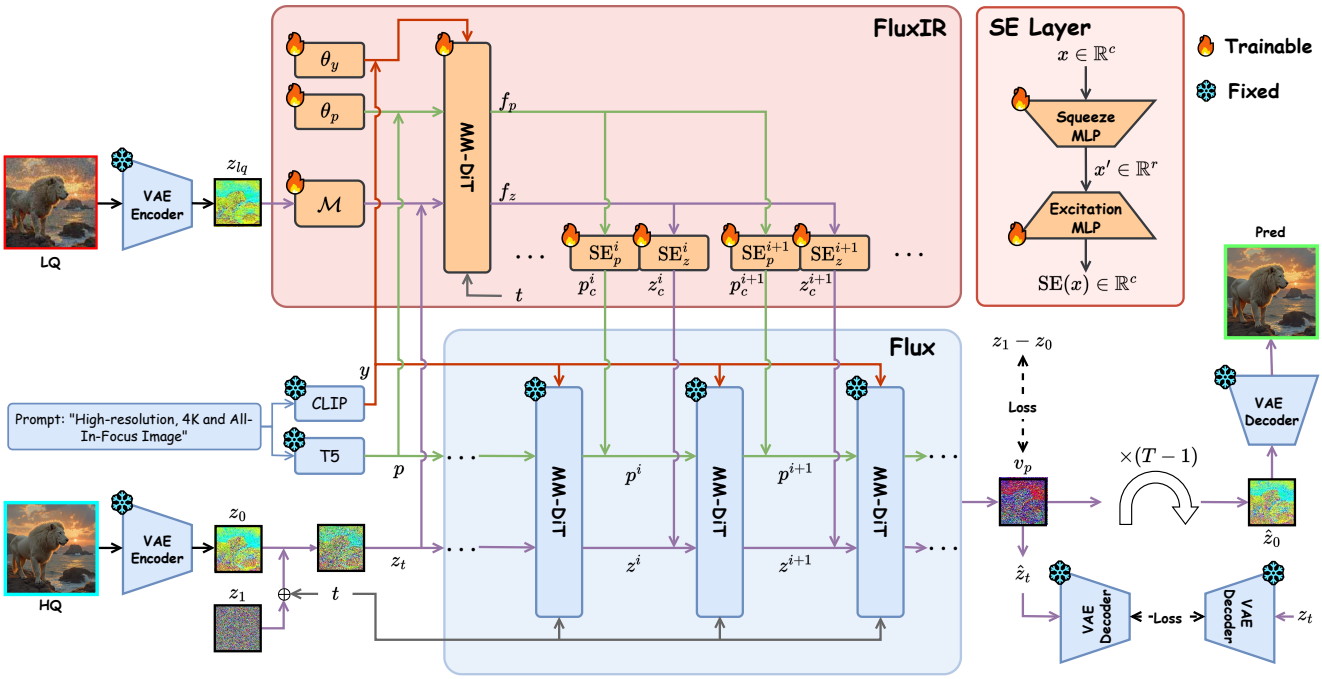

Figure 3. Training and inference pipeline of the proposed FluxIR. FluxIR employs a single MM-DiT block with learnable T5 embedding θ_p and CLIP embedding θ_y to extract image feature f_z and text feature f_p from the low-quality control latent z_{lq}. The squeeze-and-excitation (SE) layers ($\text{SE}_z(\cdot)$ for image and $\text{SE}_p(\cdot)$ for text) broadcast these features to all Flux MM-DiT blocks to enable precise and multi-modality control.

animals, plants, and the moon) or particular domains (*e.g.*, nighttime scenes and aerial photography).

Image Selection. Although Flux is a powerful text-to-image model, the stability of the images it generates is not guaranteed. Low-quality images can significantly affect the training of image restoration models. To address this, we distill the dataset using various non-reference image quality assessment (IQA) models, including CLIP-IQA [72], MANIQA [88], and MUSIQ [31]. Specifically, we select images with CLIP-IQA, MANIQA, and MUSIQ scores within the top 95% as our training set. This process ensures that low-quality and failed generated images are removed from the training dataset, thus significantly enhancing our image restoration performance.

Pair-data Construction. The degradation of ground truth images has been widely studied in [75, 99], with common degradation types including blur, downsampling, noise, JPEG compression and *etc*. In this paper, we apply the synthetic degradation method [75] with the same settings as [83] to construct paired data for training.

3.2. FluxIR Architecture

A typical solution for the T2I-based image restoration is training a ControlNet [100] upon a T2I model, with a VAE encoder to project the input image into latent space. The ControlNet is initialized as a copy of the U-Net encoder, with zero convolution layers acting as bridges to integrate conditional controls. Then the extracted multi-level features are injected into corresponding layers of the U-Net decoder. Recently, DreamClear [3] introduced a DiT-based ControlNet model by duplicating all 28 DiT blocks from PixArt-α [9]. With 2.2B trainable parameters in ControlNet, it requires 224 GPU days for training, which is surprisingly time-consuming. To solve this issue, we carefully design a training-friendly 0.4B adapter, integrated with Flux — a much larger MM-DiT-based text-to-image model with approximately 12B parameters.

The detailed FluxIR architecture is illustrated in Fig. 3. To extract control signal from VAE embedding feature z_{lq}, only one MM-DiT block is involved in our FluxIR adapter for lightweight purposes. The MM-DiT block, denoted as $\mathcal{D}(z, p, y, t; \Theta_d)$, is initialized from the first MM-DiT block of the pre-trained Flux model, and its input can be reformulated as

$$z_c = z_t + \mathcal{M}_i(z_{lq}; \Theta_i), \tag{1}$$

where $\mathcal{M}_i(\cdot; \Theta_i)$ is a multi-layer perceptron (MLP) with parameters Θ_i initialized to zero and $z_t = (1-t)z_0 + tz_1$ represents the linear combination of the noise latent z_1 and the ground truth latent z_0 at timestep t. To bridge the gap between the adapter and original T2I model, we employ a learnable T5 [56] embedding θ_p and a learnable CLIP [55]

embedding θ_y and adjust the T5 embedding p as $p_c = p+\theta_p$ and the CLIP embedding y as $y_c = y+\theta_y$. Finally, the MM-DiT block with parameters Θ_d outputs an image feature f_z and a text feature f_p as follows:

$$f_z, f_p = \mathcal{D}(z_c, p_c, y_c, t; \Theta_d). \tag{2}$$

Since Flux consists of 57 MM-DiT blocks, duplicating conditional features from a single MM-DiT adapter is insufficient to control the entire T2I model. We introduce a set of squeeze-and-excitation (SE) layers to broadcast control signals to all 57 MM-DiT blocks. Each SE layer selectively emphasizes informative features and suppresses less useful ones, enabling targeted control of its corresponding MM-DiT block. The SE layer comprises a squeeze MLP layer and an excitation MLP layer, formulated as follows:

$$\mathrm{SE}(x) = \mathbf{W}_e(\mathbf{W}_s x + \mathbf{B}_s) + \mathbf{B}_e, \tag{3}$$

where $\mathbf{W}_s \in \mathbb{R}^{r\times c}$ and $\mathbf{B}_s \in \mathbb{R}^r$ is the weight and bias of squeeze MLP layers with input channel of c and rank of r, while the weight $\mathbf{W}_e \in \mathbb{R}^{c\times r}$ and bias $\mathbf{B}_e \in \mathbb{R}^c$ of excitation layer are initialized to zero values. Unlike existing works [3, 44, 73, 91] that control only the image branch, we implement multi-modality controls on both the image and text information. The control signals of the i-th SE layer, corresponding to the i-th Flux MM-DiT block, can be computed as

$$z_c^i = \mathrm{SE}_z^i(f_z),\ p_c^i = \mathrm{SE}_p^i(f_p). \tag{4}$$

Compared with a full-rank MLP, our proposed SE layer is extremely lightweight with only approximately 2% of the parameters. Our SE layer shares the same name as SE block in [28], but differs in both motivation and implementation. [28] squeezes spatial features to extract channel-level attention for recalibrating the original feature, while our SE layers reduce the channel dimension and distill the essential features for targeted control.

3.3. Efficient Training Strategy

Timestep Sampling. The timestep sampling strategy can improve the training process for diffusion model [90]. Stable Diffusion 3 [18] proposes a logit-normal sampling strategy to emphasize training velocity when t is in the middle of $[0, 1]$. However, directly employing a logit-normal sampling strategy for IR model training is inappropriate. The starting point of the inference process is pure Gaussian noise at $t = 1$, which is seldom sampled during training [43]. This leads to ineffective IR control at $t = 1$, negatively affecting subsequent iterations. Therefore, we rewrite the timestep sampling function to ensure accurate control at the starting point. The sampling function is defined as follows:

$$t = f(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u > 1 \\ u & \text{otherwise,} \end{cases} \tag{5}$$

$$u \sim \mathcal{U}(-\epsilon, 1+\epsilon), \tag{6}$$

where the hyper-parameter ϵ set to 0.05 in our experiments. Here we first sample the temporary timestep u from a uniform distribution within the range of $(-\epsilon, 1+\epsilon)$. And then we clamp u to the range $[0, 1]$, resulting in a probability of $\frac{\epsilon}{1+2\epsilon}$ for sampling the values 0 and 1, respectively.

Optimization Strategy. The typical rectified flow model [4, 45, 46] is trained to predict the velocity field v_p, and the Mean Squared Error (MSE) loss is employed for supervising the error between predicted velocity field v_p and ground truth velocity field v_{gt} as

$$v_{gt} = z_1 - z_0, \tag{7}$$

$$\mathcal{L}_{\text{MSE}} = \|v_p - v_{gt}\|_2^2. \tag{8}$$

However, this loss function in latent space inevitably ignores the high-frequency information of image [93], which is crucial for image restoration tasks. Following [32, 33, 81, 89, 93], we incorporate a pixel-space loss function to supervise the error between the decoded latent $\hat{z}_t = z_0 + t \cdot v_p$ and the decoded ground truth latent $z_t = z_0 + t \cdot v_{gt}$ after applying the VAE decoder $\mathcal{V}_d$ into pixel space:

$$\mathcal{L}_P = \|\mathcal{V}_d(z_0 + t \cdot v_p) - \mathcal{V}_d(z_0 + t \cdot v_{gt})\|_1. \tag{9}$$

The final loss function of FluxIR is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_P, \tag{10}$$

where α is set to 1 in our experiment setting.

4. Experiments

4.1. Experimental Settings

Test Datasets. Following [3, 44, 73, 83, 91], we evaluate our method on both synthetic and real-world datasets. For the synthetic one, we randomly crop 600 images from the validation dataset of DIV2K with size of 1024×1024 and degrade them using the same settings as training, named *DIV2K-Val*. For real-world datasets, we utilize the most commonly used *RealSR* [5] and *DrealSR* [78] datasets, center-cropping HQ images to 1024×1024 and LQ images to 256×256. Additionally, we include *RealLQ250* from [3, 83, 91], a dataset of 250 LQ images at 256×256 resolution without corresponding HQ images.

Implementation Details. In the FluxGen pipeline, we generate images at a resolution of 1024×768, using the

Table 1. Quantitative comparison against state-of-the-art methods of image restoration on both synthetic and real-world datasets. The best and the second best performance for each metric are highlighted in **red** and **blue**, respectively.

Datasets	Metrics	Methods											
		BSRGAN	Real-ESRGAN	SwinIR	DASR	StableSR	DiffBIR	ResShift	SinSR	SeeSR	SUPIR	Dream-Clear	Ours
DIV2K-Val	CLIPIQA ↑	0.4741	0.4695	0.4513	0.3948	0.4085	0.5640	0.4240	0.4915	**0.5909**	0.5285	0.4961	**0.5934**
	MUSIQ ↑	63.6223	63.8292	63.6854	59.4515	55.9995	69.0996	60.4907	64.4171	**70.8168**	68.8143	67.1506	**69.6956**
	MANIQA ↑	0.4717	0.5197	0.5211	0.4316	0.4961	0.5978	0.4950	0.5059	0.5909	**0.6043**	0.5941	**0.6331**
	PSNR ↑	**21.6118**	21.5077	21.0653	**21.5082**	20.0416	21.3925	21.4906	20.7884	21.2923	20.1696	19.7605	19.3811
	SSIM ↑	0.5742	**0.5821**	**0.5762**	0.5674	0.5318	0.5297	0.5557	0.5015	0.5603	0.5145	0.4963	0.4574
	LPIPS ↓	0.3366	0.3045	0.3115	0.3320	0.3459	0.3172	0.3086	0.3443	**0.2799**	0.3184	**0.2895**	0.3888
RealSR	CLIPIQA ↑	0.4378	0.4197	0.4020	0.3153	0.4318	0.5310	0.3874	0.4445	**0.5513**	0.4594	0.5094	**0.5413**
	MUSIQ ↑	63.4256	60.9526	58.6029	45.7800	58.8339	65.5924	54.4401	59.4758	**69.0118**	63.9510	64.1779	**67.4536**
	MANIQA ↑	0.5059	0.5277	0.4765	0.3875	0.5469	0.5888	0.4688	0.5232	**0.6115**	0.5903	0.5997	**0.6334**
	PSNR ↑	**24.6354**	24.1773	24.4184	**25.0954**	20.4413	23.9748	24.1925	24.2417	24.1676	22.3887	22.0330	20.7176
	SSIM ↑	0.7484	0.7482	**0.7566**	**0.7569**	0.6320	0.6766	0.6892	0.6637	0.6967	0.6450	0.6442	0.5269
	LPIPS ↓	0.2109	**0.2082**	**0.2049**	0.2495	0.2305	0.2484	0.2440	0.2601	0.2105	0.2674	0.2547	0.3672
DrealSR	CLIPIQA ↑	0.4219	0.3922	0.3878	0.3165	0.4206	0.4889	0.4060	0.4188	**0.5295**	0.4544	0.4046	**0.5136**
	MUSIQ ↑	61.2255	58.3819	57.3308	46.4866	56.3195	62.0861	53.9293	58.3077	**67.2415**	64.7508	56.5999	**66.6202**
	MANIQA ↑	0.4823	0.4911	0.4710	0.3828	0.5224	0.5566	0.4611	0.4830	**0.5928**	0.5764	0.5423	**0.6024**
	PSNR ↑	24.0480	24.1436	23.8878	**25.1821**	21.9354	23.9912	22.9158	23.4737	**24.1699**	22.6023	22.8037	21.3549
	SSIM ↑	0.7268	**0.7390**	0.7290	**0.7689**	0.6616	0.6364	0.6330	0.6190	0.7147	0.6406	0.6186	0.5675
	LPIPS ↓	0.2257	**0.2226**	**0.2175**	0.2474	0.2357	0.3123	0.3263	0.3544	0.2396	0.3069	0.2905	0.4310
RealLQ250	CLIPIQA ↑	0.4701	0.4359	0.4400	0.3486	0.4009	0.5377	0.4165	0.4810	**0.5569**	0.4808	0.4876	**0.5639**
	MUSIQ ↑	63.5206	62.5161	63.3724	53.0238	56.7121	67.5326	59.5056	63.8644	**70.3768**	65.7804	66.5102	**70.7770**
	MANIQA ↑	0.5007	0.5239	0.5335	0.4414	0.5144	0.5877	0.5005	0.5160	**0.5927**	0.5829	0.5853	**0.6314**

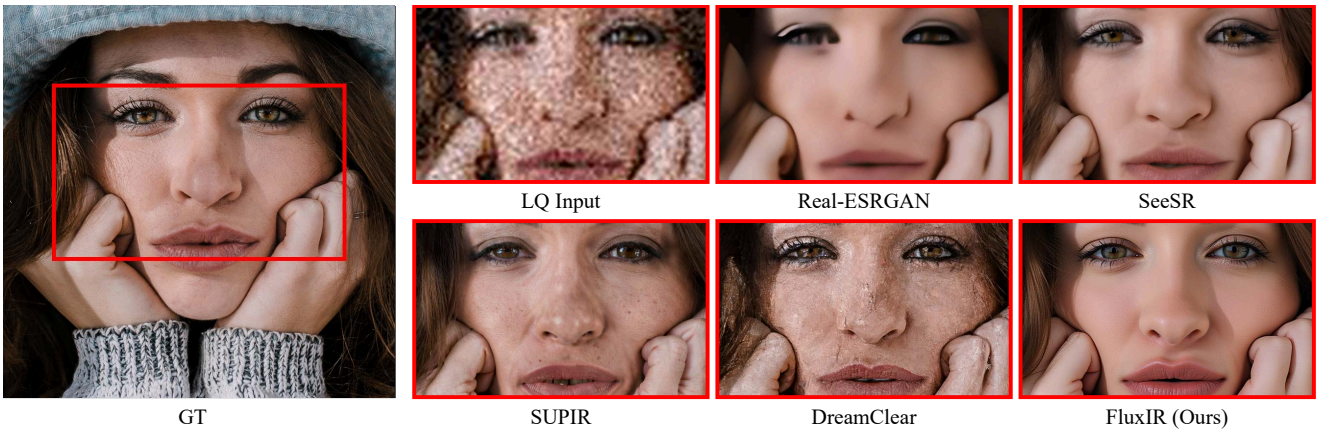

Figure 4. Qualitative comparison on the synthetic dataset *DIV2K-Val*.

guidance scale of 4 and timesteps of 20. In all experiments, we exclusively use 350,000 images generated by FluxGen, with no external data involved. We randomly center-crop the images to the sizes of 1024×768, 768×768, and 512×768 to accommodate different aspect ratios. We train our FluxIR model using the AdamW [48] optimizer with a learning rate of 1×10^{-5}. The training process takes about 3.5 days on 4 NVIDIA H800 GPUs with a batch size of 64. For inference, we adopt 20 sampling steps to generate our IR results across all of our experiments.

Metrics. Following [91], we use PSNR, SSIM, and LPIPS as full-reference metrics and MANIQA, CLIPIQA, and MUSIQ as non-reference metrics[2] for comprehensive evaluation. Our method, like other generative IR approaches [3, 44, 73, 83, 91], achieves strong results in non-reference metrics but performs less competitively on full-reference metrics.

4.2. Comparison with Existing Methods

We conduct both quantitative and qualitative comparisons of our FluxIR with state-of-the-art image restoration methods, including early approaches (BSRGAN [99],

[2]We utilize the GitHub repository - https://github.com/chaofengc/IQA-PyTorch to evaluate non-reference metric scores with model names "maniqa-pipal", "clipiqa+_vitL14_512", and "musiq", respectively.

Figure 5. Qualitative comparison on the real-world dataset *RealLQ250*.

RealESRGAN[75], SwinIR [40] and DASR[41]) and recent generative IR models (StableSR [73], DiffBIR [44], ResShift [92], SinSR [76], SeeSR [83], SUPIR [91], DreamClear [3]).

Quantitative Comparisons. Tab. 1 shows the quantitative comparisons with the existing SOTAs on the four datasets. Our FluxIR model achieves the best score in MANIQA with a significant margin across all synthetic and real-world datasets. On the *RealLQ250* dataset, our method significantly outperforms other methods across all non-reference metrics, while MANIQA even surpasses the second-best method by over 6.53%. On the other three synthetic and real-world datasets, our approach achieves the best or second-best scores in all non-reference metrics (CILPIQA, MUSIQ, and MANIQA). It is worth mentioning that our proposed method trained within 14 GPU-days significantly reduces training costs by approximately 93.5% and 91.5% compared to SUPIR [91] and DreamClear [3], respectively.

Qualitative Comparisons. We present a qualitative comparison on the synthetic dataset *DIV2K-Val* in Fig. 4. Our method demonstrates a notable advantage in generating high-frequency details, such as lip texture, hair, eyelashes, and pupils in portrait images. In contrast, SUPIR [91] and DreamClear [3] fail to accurately reproduce human skin textures, and SeeSR [83] produces overly smooth results, lacking the natural skin texture and fine details. For the real-world dataset *RealLQ250*, we provide a qualitative comparison in Fig. 5. In the first row, which shows restoration results for a low-quality image of a cat, our method effectively reconstructs realistic high-frequency details such as fur, whiskers, and eyes. SeeSR [83] restores some fur and whiskers but lacks overall realism, while other methods fail to produce a high-quality, realistic result. In the second row, focusing on text recovery, our method achieves the most accurate text reconstruction, whereas other methods introduce various text distortions. Specifically, SUPIR [91] produces blurry results, and SeeSR [83] mistakenly recovers incorrect text content.

4.3. Ablation Study

Effectiveness of SE layer. We evaluate the impact of different ranks of our SE layer by setting r to 16, 32, 64, and 128, respectively. Another variant is replacing the SE layer with a single MLP, *i.e.* a full rank setting. As shown in Tab. 2, the SE layer with $r = 32$ achieves the best performance on CLIPIQA, MANIQA, and CLIPIQA. Compared to using full-rank MLPs, the SE layer with low-rank results in a comparable performance while saving over 72% of

Table 2. Ablation results of various SE layer ranks on the *RealLQ250* dataset. The number of parameters in FluxIR without the SE layers is 387M.

Rank	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑	# Params
16	0.5389	69.72	0.6222	387M + 11.6M
32	0.5639	70.78	0.6314	387M + 22.8M
64	0.5529	70.37	0.6306	387M + 45.2M
128	0.5606	70.54	0.6292	387M + 90.0M
Full	0.5720	70.79	0.6324	387M + 1076.2M

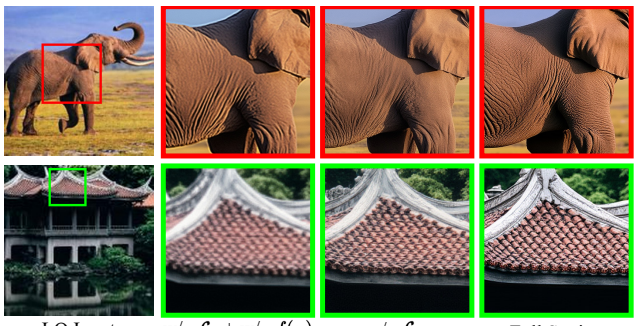

Figure 6. The visual comparison results of our training strategies: timestep sampling function $f(u)$ and $\mathcal{L}_P$ loss. Please zoom in for a better view.

the parameters. This demonstrates the effectiveness of our designed light-weighted SE layers in controlling the Flux model. By utilizing our SE layers with a rank of 32, we can outperform other state-of-the-art (SOTA) methods using only 0.4B adapter parameters, which is only 30.8% and 18.6% the size of SUPIR and DreamClear, respectively.

Effectiveness of Training Strategies. We conduct ablation studies to evaluate our proposed training strategies: timestep sampling function $f(u)$ and $\mathcal{L}_P$ loss. Firstly, we use the logit-normal sampling strategy [18] to train our model without $\mathcal{L}_P$ loss as a baseline. Then we include our proposed timestep sampling function $f(u)$ as the second setting. We compare these two experiments with our full training strategies in the Tab. 3, which clearly demonstrates the significant performance improvement of the new sampling function and $\mathcal{L}_P$ loss. Fig. 6 illustrates that the logit-normal sampling produces low-quality content, while our proposed sampling function $f(u)$ effectively addresses this issue by bridging the gap between training and inference. Additionally, applying the pixel-space loss $\mathcal{L}_P$ allows for the restoration of more high-frequency details.

Table 3. Ablation comparison results of our training strategies on the *RealLQ250* dataset.

Strategy	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑
w/o $\mathcal{L}_P$ and w/o $f(u)$	0.4416	62.18	0.5720
w/o $\mathcal{L}_P$	0.5082	68.53	0.6128
Full Setting	**0.5639**	**70.78**	**0.6314**

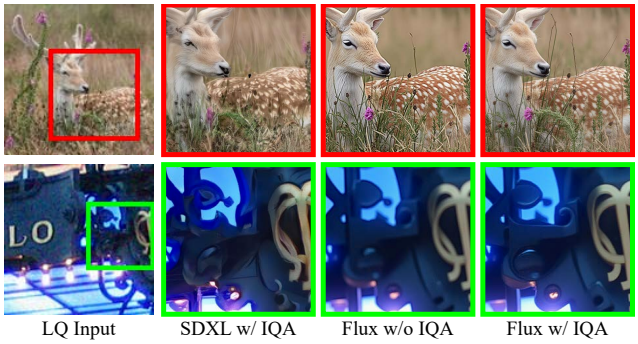

Figure 7. The visual comparisons of different FluxGen settings, where we study different T2I models, *i.e.* SDXL and Flux, and the usage of IQA selections. Please zoom in for a better view.

Effectiveness of FluxGen. We follow the same settings of our FluxGen pipeline but replace the T2I model from Flux with SDXL to generate $10,000$ images. Additionally, we remove the IQA-based selection from the FluxGen pipeline. The results in Tab. 4 demonstrate that a high-quality image dataset from the Flux model, combined with IQA selection, leads to high-quality IR results. Fig. 7 shows that the visual results based on the SDXL dataset are abysmal, exhibiting a smearing and distortion effect. In contrast, images generated by Flux deliver the best visual aesthetic quality. Our IQA selection can eliminate poorly generated images to achieve more realistic results. The samples of generated images by FluxGen can be found in the supplemental materials.

Table 4. Ablation results of different datasets generated by three FluxGen settings on the *RealLQ250* dataset.

FluxGen	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑
SDXL w/ IQA	**0.5787**	69.28	0.5630
Flux w/o IQA	0.5254	68.66	0.6265
Flux w/ IQA	0.5487	**70.16**	**0.6267**

5. Conclusion

In this paper, we squeezed out the powerful T2I model - Flux for image restoration, by acquiring training data from it and then building a lightweight adapter to control it. To tackle the challenge of acquiring a large-scale high-quality image dataset, we proposed FluxGen for streamlined and highly efficient data generation. We present FluxIR, a lightweight adapter designed to control the T2I model for real-world image restoration, where the SE layers broadcast the control signals to all Flux MM-DiT blocks and modulate both image and text embedding to enable precise and multi-modality control. Extensive qualitative and quantitative experiments demonstrated that our method shows superior performance on image restoration tasks and significantly reduces the training cost of large-scale generative IR model, in comparison with previous state-of-the-arts.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 2

[3] 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. *arXiv preprint arXiv:2410.18666*, 2024. 1, 2, 3, 4, 5, 6, 7

[4] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 5

[5] 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 *ICCV*, pages 3086–3095, 2019. 5

[6] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *ACM MM*, pages 1329–1338, 2022. 2

[7] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *CVPR*, pages 1692–1703, 2023. 1

[8] Haolan Chen, Jinhua Hao, Kai Zhao, Kun Yuan, Ming Sun, Chao Zhou, and Wei Hu. Cassr: Activating image power for real-world image super-resolution. *arXiv preprint arXiv:2403.11451*, 2024. 2

[9] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 2, 4

[10] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, pages 12312–12321, 2023. 1

[11] Zheng Chen, Yulun Zhang, Jinjin Gu, Xin Yuan, Linghe Kong, Guihai Chen, and Xiaokang Yang. Image super-resolution with text prompt diffusion. *arXiv preprint arXiv:2311.14282*, 2023. 2

[12] Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *NeurIPS*, 36, 2024. 1

[13] Qinpeng Cui, Yixuan Liu, Xinyi Zhang, Qiqi Bao, Zhongdao Wang, Qingmin Liao, Li Wang, Tian Lu, and Emad Barsoum. Taming diffusion prior for image super-resolution with domain shift sdes. *arXiv preprint arXiv:2409.17778*, 2024. 2

[14] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 2

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 2

[16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014. 2

[17] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 1

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2, 3, 5, 8

[19] Yuanting Fan, Chengxu Liu, Nengzhong Yin, Changlong Gao, and Xueming Qian. Adadiffsr: Adaptive region-aware dynamic acceleration diffusion model for real-world image super-resolution. *arXiv preprint arXiv:2410.17752*, 2024. 2

[20] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCVW*, pages 3599–3608. IEEE, 2019. 2

[21] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[22] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML Workshop*, 2023. 3

[23] Junhao Gu, Peng-Tao Jiang, Hao Zhang, Mi Zhou, Jinwei Chen, Wenming Yang, and Bo Li. Consissr: Delving deep into consistency in diffusion-based image super-resolution. *arXiv preprint arXiv:2410.13807*, 2024. 2

[24] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *ICCVW*, pages 3512–3516. IEEE, 2019. 2

[25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[26] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *CVPR*, pages 1732–1741, 2019. 2

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[28] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 5

[29] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, pages 1689–1697, 2017. 1, 2

[30] Jen-Yuan Huang, Haofan Wang, Qixun Wang, Xu Bai, Hao Ai, Peng Xing, and Jen-Tse Huang. Instantir: Blind image restoration with instant generative reference. *arXiv preprint arXiv:2410.06551*, 2024. 2

[31] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 2, 4

[32] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2426–2435, 2022. 5

[33] Jinseok Kim and Tae-Kyun Kim. Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder. In *CVPR*, pages 9202–9211, 2024. 5

[34] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 1–21. IEEE, 2017. 1

[35] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2013. 3

[36] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3

[37] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 1, 2

[38] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Distillation-free one-step diffusion for real-world image super-resolution. *arXiv preprint arXiv:2410.04224*, 2024. 2

[39] 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 *CVPR*, pages 1775–1787, 2023. 2

[40] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 2, 7

[41] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, pages 574–591. Springer, 2022. 2, 7

[42] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 2

[43] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, pages 5404–5411, 2024. 5

[44] 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. 2, 5, 6, 7

[45] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5

[46] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5

[47] Yong Liu, Hang Dong, Jinshan Pan, Qingji Dong, Kai Chen, Rongxiang Zhang, Lean Fu, and Fei Wang. Patcscaler: An efficient patch-independent diffusion model for super-resolution. *arXiv preprint arXiv:2405.17158*, 2024. 2

[48] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[49] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, pages 291–300, 2020. 2

[50] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2, 3

[51] 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. 2, 3

[52] Chenyang Qi, Zhengzhong Tu, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Qifeng Chen, and Hossein Talebi. Tip: Text-driven image processing with semantic and restoration instructions. *arXiv preprint arXiv:2312.11595*, 2023. 2

[53] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *ECCV*, pages 285–303. Springer, 2025. 2

[54] Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *CVPR*, pages 5754–5763, 2023. 1

[55] 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 *ICML*, pages 8748–8763. PMLR, 2021. 4

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3, 4

[57] 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. 3

[58] Yulin Ren, Xin Li, Bingchen Li, Xingrui Wang, Mengxi Guo, Shijie Zhao, Li Zhang, and Zhibo Chen. Moe-diffir: Task-customized diffusion priors for universal compressed image restoration. In *ECCV*, pages 116–134. Springer, 2025. 3

[59] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201. Springer, 2020. 1

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3

[61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2, 3

[62] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, pages 4491–4500, 2017. 2

[63] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 2

[64] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32, 2019. 2

[65] 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 *CVPR*, pages 25868–25878, 2024. 2

[66] Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability of diffusion models for content consistent super-resolution. *arXiv e-prints*, pages arXiv–2401, 2023. 2

[67] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 1

[68] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[69] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[70] Yuhao Wan, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, Ming-Ming Cheng, and Bo Li. Clearsr: Latent low-resolution image embeddings help diffusion-based real-world super resolution models see clearer. *arXiv preprint arXiv:2410.14279*, 2024. 2

[71] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, pages 2747–2757, 2020. 2

[72] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, pages 2555–2563, 2023. 2, 4

[73] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, pages 1–21, 2024. 2, 5, 6, 7

[74] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, pages 10581–10590, 2021. 2

[75] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 2, 4, 7

[76] 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 *CVPR*, pages 25796–25805, 2024. 7

[77] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *CVPR*, pages 18156–18165, 2023. 1

[78] 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 *ECCV*, pages 101–117. Springer, 2020. 5

[79] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*, pages 13385–13394, 2021. 2

[80] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *CVPR*, pages 16293–16303, 2022. 1

[81] Hanlin Wu, Jiangwei Mo, Xiaohui Sun, and Jie Ma. Latent diffusion, implicit amplification: Efficient continuous-scale super-resolution for remote sensing images. *arXiv preprint arXiv:2410.22830*, 2024. 5

[82] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 2

[83] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pages 25456–25467, 2024. 2, 4, 5, 6, 7

[84] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE TPAMI*, 45(3):3121–3138, 2022. 3

[85] Qianlong Xiang, Miao Zhang, Yuzhang Shang, Jianlong Wu, Yan Yan, and Liqiang Nie. Dkdm: Data-free knowledge distillation for diffusion models with any architecture. *arXiv preprint arXiv:2409.03550*, 2024. 3

[86] Rui Xie, Ying Tai, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, Xiaoqian Ye, Qian Wang, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717*, 2024. 2

[87] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou,

Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3

[88] 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 *CVPR*, pages 1191–1200, 2022. 2, 4

[89] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 2, 5

[90] Jingfeng Yao, Wang Cheng, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *arXiv preprint arXiv:2410.10356*, 2024. 5

[91] 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*, pages 25669–25680, 2024. 1, 2, 3, 5, 6, 7

[92] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 36, 2024. 7

[93] Christina Zhang, Simran Motwani, Matthew Yu, Ji Hou, Felix Juefei-Xu, Sam Tsai, Peter Vajda, Zijian He, and Jialiang Wang. Pixel-space post-training of latent diffusion models. *arXiv preprint arXiv:2409.17565*, 2024. 5

[94] Haiyu Zhang, Shaolin Su, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Gsdd: Generative space dataset distillation for image super-resolution. In *AAAI*, pages 7069–7077, 2024. 3

[95] Jiahui Zhang, Shijian Lu, Fangneng Zhan, and Yingchen Yu. Blind image super-resolution via contrastive representation learning. *arXiv preprint arXiv:2107.00708*, 2021. 2

[96] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 1

[97] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE TIP*, 27(9):4608–4622, 2018. 1

[98] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, pages 2737–2746, 2020. 1

[99] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 2, 4, 6

[100] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 4

[101] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 2

[102] Yuhong Zhang, Hengsheng Zhang, Xinning Chai, Zhengxue Cheng, Rong Xie, Li Song, and Wenjun Zhang. Diff-restorer: Unleashing visual prompts for diffusion-based universal image restoration. *arXiv preprint arXiv:2407.03636*, 2024. 3

[103] Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow. In *CVPR*, pages 13–22, 2024. 2