Photo-Realistic Image Restoration in the Wild with Controlled Vision-Language Models

Ziwei Luo[1] Fredrik K. Gustafsson[2] Zheng Zhao[1] Jens Sjölund[1] Thomas B. Schön[1]
[1]Uppsala University [2]Karolinska Institutet
{ziwei.luo,zheng.zhao,jens.sjolund,thomas.schon}@it.uu.se, fredrik.gustafsson@ki.se
https://github.com/Algolzw/daclip-uir

Synthetic LQ | Real-ESRGAN | StableSR | SUPIR | Our result

Figure 1. Examples of the synthetic LQ images generated using our proposed degradation pipeline and results produced by our method and other state-of-the-art wild IR approaches: Real-ESRGAN [53], StableSR [51], and SUPIR [57]. Notably, both StableSR and SUPIR adapt pretrained Stable Diffusion [39, 44] models to image restoration, and SUPIR further leverages textual semantic guidance using LLaVA [26]. The proposed method successfully handles various complex degradations and produces clean and sharp results.

Abstract

Though diffusion models have been successfully applied to various image restoration (IR) tasks, their performance is sensitive to the choice of training datasets. Typically, diffusion models trained in specific datasets fail to recover images that have out-of-distribution degradations. To address this problem, this work leverages a capable vision-language model and a synthetic degradation pipeline to learn image restoration in the wild (wild IR). More specifically, all low-quality images are simulated with a synthetic degradation pipeline that contains multiple common degradations such as blur, resize, noise, and JPEG compression. Then we introduce robust training for a degradation-aware CLIP model to extract enriched image content features to assist high-quality image restoration. Our base diffusion model is the image restoration SDE (IR-SDE). Built upon it, we further present a posterior sampling strategy for fast noise-free image generation. We evaluate our model on both synthetic and real-world degradation datasets. Moreover, experiments on the unified image restoration task illustrate that the proposed posterior sampling improves image generation quality for various degradations.

1. Introduction

Diffusion models have proven effective for high-quality (HQ) image generation in various image restoration (IR) tasks such as image denoising [7, 14, 30], deblurring [6, 43, 55], deraining [37, 59], dehazing [30, 31], inpainting [28, 44, 47], super-resolution [18, 48, 60], shadow removal [11, 31], etc. Compared to traditional deep learning-based approaches that directly learn IR models using an ℓ_1 or ℓ_2 loss [5, 23, 61, 62] or an adversarial loss [16, 52, 53], diffusion models are known for their ability to generate photo-realistic images with a stable training process. However, they are mostly trained on fixed datasets and therefore typically fail to recover high-quality outputs when applied to real-world scenarios with unknown, complex, out-of-distribution degradations [53].

Although this problem can be alleviated by leveraging large-scale pretrained *Stable Diffusion* [39, 44] weights [25, 51, 57] and synthetic low-quality (LQ) image generation pipelines [46, 53], it is still challenging to accurately restore real-world images in the wild (i.e., *wild IR*). On the one hand, Stable Diffusion uses an adversarially trained variational autoencoder (VAE) to compress the diffusion to latent space, which is efficient but loses image details in the reconstruction process. Moreover, in practice, the restoration in latent space is unstable and tends to generate color-shifted images [25]. On the other hand, most existing works use a fixed degradation pipeline (with different probabilities for each degradation) to generate low-quality images [53], which might be insufficient to represent the complex real-world degradations.

In this work, we aim to perform photo-realistic image restoration with enriched vision-language features that are extracted from a degradation-aware CLIP model (DA-CLIP [32]). For scenes encountered in the wild, we assume the image only contains mild, common degradations such as light noise and blur, which can be difficult to represent by text descriptions. We thus add a fidelity loss to reduce the distance between the LQ and HQ image embeddings. Then the enhanced LQ embedding is incorporated into the image restoration networks (such as the U-Net [45] in IR-SDE [30]) via cross-attention. Inspired by Real-ESRGAN [53], we also propose a new degradation pipeline with a random shuffle strategy to improve the generalization. An optimal posterior sampling strategy is further proposed for IR-SDE to improve its performance. Fig. 1 shows the comparison of our method with other state-of-the-art wild IR approaches.

In summary, our main contributions are as follows:

- We present a new synthetic image generation pipeline that employs a random shuffle strategy to simulate complex real-world LQ images.
- For degradations in the wild, we modify DACLIP to reduce the embedding distance of LQ-HQ pairs, which enhances LQ features with high-quality information.
- We propose a posterior sampling strategy for IR-SDE [30] and show that it is the optimal reverse-time path, yielding a better image restoration performance.
- Extensive experiments on wild IR and other specific IR tasks demonstrate the effectiveness of each component of our method.

2. Related Work

Blind Image Restoration Image restoration (IR) aims to reconstruct a high-quality (HQ) image from its corrupted counterpart, i.e. from a low-quality (LQ) image with task-specific degradations [9, 20, 22, 61–64, 69]. Most learning-based approaches directly train neural networks with an ℓ_1/ℓ_2 loss on HQ-LQ image pairs, which is effective but often overfit on specific degradations [53, 57, 65]. Thus the blind IR approach is proposed and has gained growing attention in addressing complex real-world degradations. BSRGAN [65] is the pioneering work that designs a practical degradation model for blind super-resolution, and Real-ESRGAN [53] improves it by exploiting a 'high-order' degradation pipeline. Most subsequent blind IR methods [4, 25] follow their degradation settings but with some improvements in architectures and loss functions. Recently, some works [17, 32, 40] further propose to jointly learn different IR tasks using a single model to improve the task generalization, so-called unified image restoration.

Photo-Realistic Image Restoration Starting from ESRGAN [16], photo-realistic IR becomes prevalent due to the increasing requirement for high-quality image generation. Early research explored a variety of methods that combine GANs [10, 34] and other perceptual losses [8, 13, 67] to train networks to predict images following the natural image distribution [16, 52, 53]. However, GAN-based approaches often suffer from unstable performance and can be challenging to train on small datasets. Recent works therefore introduce diffusion models in image restoration for realistic image generation [14, 18, 30, 31, 37, 48]. Moreover, leveraging pretrained Stable Diffusion (SD) models [39, 44] as the prior is growing popular in real-world and blind IR tasks [25, 51, 56, 57]. In particular, StableSR [51] and DiffBIR [25] adapt the SD model to image restoration using an approach similar to ControlNet [66]. CoSeR [50], SeeSR [56], and SUPIR [57] further introduce the textual semantic guidance in diffusion models for more accurate restoration performance.

3. Method

Our work is a set of extensions and improvements on the degradation-aware CLIP [32] which, in turn, builds on a mean-reverting SDE [30]. Thus, before going into our contributions in the following sections, we first summarize the main constructions of the mean-reverting SDE and degradation-aware CLIP.

3.1. Preliminaries

Mean-Reverting SDE Given a random variable x_0 sampled from an unknown distribution, $x_0 \sim p_0(x)$, the mean-reverting SDE [30] is defined according to

$$dx = \theta_t \, (\mu - x) dt + \sigma_t dw, \tag{1}$$

where θ_t and σ_t are predefined time-dependent coefficients and w is a standard Wiener process. By restricting the coef-

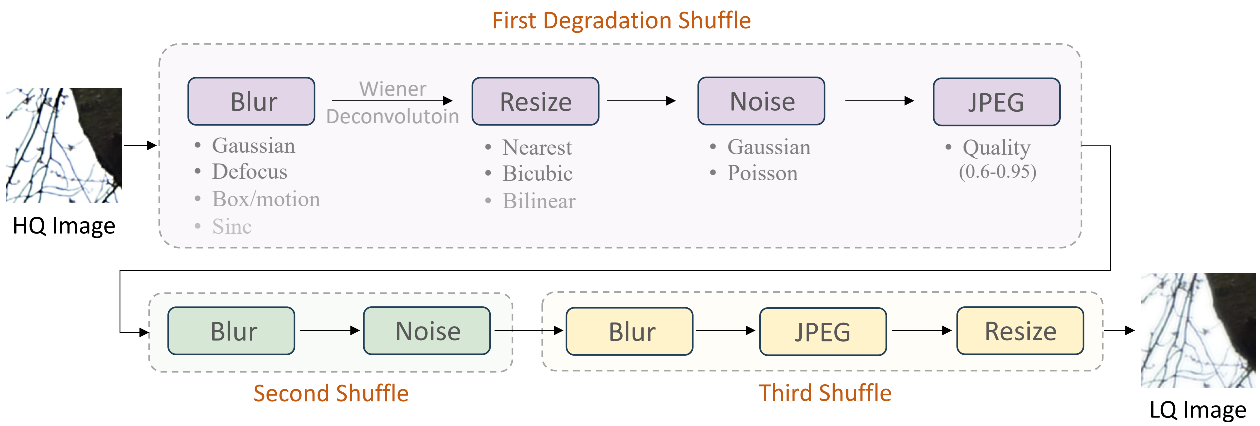

Figure 2. Overview of the proposed pipeline for synthetic image degradation. There are three degradation phases adopting the random shuffle strategy. We use different types of filters in blur generation and add the Wiener deconvolution for simulating ringing artifacts similar to the Sinc filter in Real-ESRGAN [53]. As a general $\times 1$ image restoration pipeline, we use one 'resize' operation to provide image resolution augmentation, and another resize operation to ensure that all the degraded images are resized back to their original size.

ficients to satisfy $\sigma_t^2 \,/\, \theta_t \,=\, 2\,\lambda^2$ for all timesteps t, we can solve the marginal distribution $p_t(x)$ as follows:

$$\begin{aligned} p_t(x) \ &= \mathcal{N}\big(x_t \mid m_t, v_t\big), \\ m_t &= \mu + (x_0 - \mu)\, \mathrm{e}^{-\bar{\theta}_t}, \\ v_t &= \lambda^2 \left(1 - \mathrm{e}^{-2\,\bar{\theta}_t}\right), \end{aligned} \tag{2}$$

where $\bar{\theta}_t = \int_0^t \theta_z \, \mathrm{d}z$. To simulate the image degradation process, we set the HQ image as the initial state x_0 and the LQ image as the mean μ. Then the forward SDE iteratively transforms the HQ image into the LQ image with additional noise, where the noise level is fixed to λ.

Moreover, Anderson [2] states that the forward process (Eq. (1)) has a reverse-time representation as

$$dx = \big[\theta_t\,(\mu - x) - \sigma_t^2\,\nabla_x \log p_t(x)\big]\,dt + \sigma_t d\hat{w}, \tag{3}$$

where $\nabla_x \log p_t(x)$ is the score function, which can be computed via Eq. (2) during training since we have access to the ground truth LQ-HQ pairs in the training dataset. Following IR-SDE [30], we train the score prediction network with a maximum likelihood loss which specifies the optimal reverse path x_{t-1}^* for all times:

$$\begin{aligned} x_{t-1}^* = &\ \frac{1 - \mathrm{e}^{-2\,\bar{\theta}_{t-1}}}{1 - \mathrm{e}^{-2\,\bar{\theta}_t}} \mathrm{e}^{-\theta_t^{'}} (x_t - \mu) \\ &+ \frac{1 - \mathrm{e}^{-2\,\theta_t^{'}}}{1 - \mathrm{e}^{-2\,\bar{\theta}_t}} \mathrm{e}^{-\bar{\theta}_{t-1}} (x_0 - \mu) + \mu, \end{aligned} \tag{4}$$

where $\theta_i^{'} = \int_{i-1}^{i} \theta_t dt$. The proof can be found in [30]. Once trained, we can simulate the backward SDE (Eq. (3)) to restore the HQ image, similar to what is done in other diffusion-based models [49].

Degradation-Aware CLIP The core component of the degradation-aware CLIP (DACLIP [32]) is a controller that explicitly classifies degradation types and, more importantly, adapts the fixed CLIP image encoder [42] to output high-quality content embeddings from corrupted inputs for accurate multi-task image restoration. DACLIP uses a contrastive loss to optimize the controller. Moreover, the training dataset is constructed with image-caption-degradation pairs where all captions are obtained using BLIP [19] on the clean HQ images of a multi-task dataset.

The trained DACLIP model is then applied to downstream networks to facilitate image restoration. Specifically, the cross-attention [44] mechanism is introduced to incorporate image content embeddings to learn semantic guidance from the pre-trained DACLIP. For the unified image restoration task, the predicted degradation embedding is useful and can be combined with visual prompt learning [71] modules to further improve the performance.

3.2. Synthetic Image Degradation Pipeline

To restore clean images from unknown and complex degradations, we use a synthetic image degradation pipeline for LQ image generation, as shown in Fig. 2. Common degradation models like **blur**, **resize**, **noise**, and **JPEG** compression are repeatedly involved to simulate complex scenarios. Following the *high-order degradation* in Real-ESRGAN [53], all degradation models in our pipeline have individual parameters that are randomly picked in each training step, which substantially improves the generalization for out-of-distribution datasets [29, 53, 65]. In particular, in the blur model, we add some specific filter types (e.g., defocus, box, and motion filters) rather than only Gaussian filters for more general degradations, and the Wiener deconvolution is included to simulate natural ringing artifacts (which usually occurs in the preprocessing steps in some electronic cameras [15, 58]). Wiener deconvolution generates more distinct ringing artifacts on textures than the Sinc filter [53], which can be seen in the two examples of applying Wiener deconvolution to blurry images in Fig. 3. For $\times 1$ image restoration (no resolution changes), we use two resize operations (with different interpolation modes) to provide random resolution augmentation and ensuring that all degraded images then are resized back to their original size. Note that our model focuses on image restoration in the wild (wild IR) and we therefore set all degradations to be light and diverse. Moreover, we randomly shuffle the degradation orders to further improve the generalization.

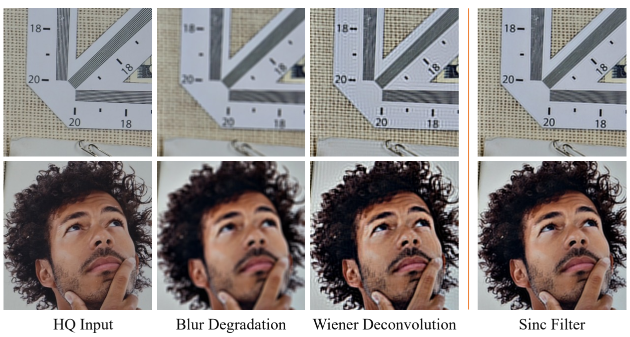

Figure 3. Examples of applying Wiener deconvolution to generate ringing artifacts. Compared to the Sinc filter used in Real-ESRGAN [53], the proposed Wiener deconvolution generates more distinct ringing artifacts on textures.

3.3. Robust Degradation-Aware CLIP

As introduced in Sec. 3.1, DACLIP leverages a large-scale pretrained vision-language model, namely CLIP, for multi-task image restoration. While it works well on some (relatively) large and distinct degradation types such as rain, snow, shadow, inpainting, etc., it fares worse on the wild IR task since most degradations are mild, hard to describe in text, and contain multiple degradations in the same image.

To address this problem, we update DACLIP to learn more robust embeddings with the following aspects: 1) In dataset construction, instead of only using one degradation for each image, we use different combinations of degradation types such as *'an image with blur, noise, ringing artifacts'* as the degradation text. 2) We add an ℓ_1 loss to minimize the embedding distance between LQ and HQ images, where the HQ image embedding is extracted from the frozen CLIP image encoder. An overview of the robust DACLIP is illustrated in Fig. 4. The multi-degradation texts enable DACLIP to handle images that contain multiple complex degradations in the wild. Moreover, the additional ℓ_1 loss forces DACLIP to learn accurate clean embeddings

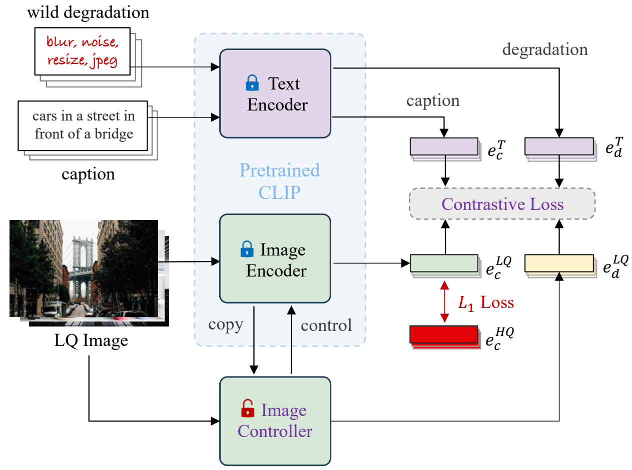

Figure 4. The proposed robust degradation-aware CLIP (DACLIP) model. e_c^T and e_d^T are caption and degradation text embeddings, respectively. The embeddings (e_c^{LQ}, e_d^{LQ}) are extracted from LQ images, and e_c^{HQ} represents the HQ image embedding extracted from the original CLIP image encoder.

from our synthetic corrupted inputs.

As Luo et al. [32] illustrates, the quality of the image content embedding significantly affects the restoration results, thus encouraging us to extend the DACLIP base encoders to larger models for better performance. Specifically, we first generate clean captions using HQ images and then train the ViT-L-14 (rather than ViT-B-32 in DACLIP) based on the synthetic image-caption-degradation pairs, where the LQ images are generated following the pipeline in Fig. 2. The dimensions of both image and text embeddings have increased from 512 to 768, which introduces more details for downstream IR models.

We use IR-SDE [30] for realistic image generation and insert the image content embedding into the U-Net via cross-attention [44], analogously to what was done in [32]. Since the degradation level is difficult to describe using text (e.g., the blurry level, noise level, and quality compression rate), we thus abandon the use of degradation embeddings for wild image restoration in both training and testing, similar to the single task setting in DACLIP [32]. In addition, to enable large-size inputs, we simply modify the network with an additional downsampling layer and an upsampling layer before and after the U-Net for model efficiency.

3.4. Optimal Posterior Sampling for IR-SDE

It is worth noting that the forward SDE in Eq. (1) requires many timesteps to converge to a relatively stable state, i.e. a noisy LQ image with noise level λ. The sampling process (HQ image generation) uses the same timesteps as the forward SDE and is also sensitive to the noise scheduler [36]. To improve the sample efficiency, Zhang et al. [68] propose a posterior sampling approach by specifying the optimal mean and variance in the reverse process. However, their method sets the SDE mean μ to 0, and only uses it to generate actions as a typical diffusion policy in reinforcement learning applications. In this work, we extend their posterior sampling strategy into a more general form for IR-SDE.

Let us use the same notation as in Sec. 3.1. Formally, given the initial state x_0 and any other diffusion state x_t at time $t \in [1, T]$, we can prove that the posterior of the mean-reverting SDE is tractable when conditioned on x_0. More specifically, this posterior distribution is given by

$$p(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}_t(x_t, x_0),\ \tilde{\beta}_t I), \quad (5)$$

which is a Gaussian with mean and variance given by:

$$\begin{aligned}\tilde{\mu}_t(x_t, x_0) = & \frac{1 - \mathrm{e}^{-2\,\bar{\theta}_{t-1}}}{1 - \mathrm{e}^{-2\,\bar{\theta}_t}} \mathrm{e}^{-\theta_t'} (x_t - \mu) \\ & + \frac{1 - \mathrm{e}^{-2\,\theta_t'}}{1 - \mathrm{e}^{-2\,\bar{\theta}_t}} \mathrm{e}^{-\bar{\theta}_{t-1}} (x_0 - \mu) + \mu,\end{aligned} \quad (6)$$

$$\text{and} \quad \tilde{\beta}_t = \frac{(1 - \mathrm{e}^{-2\bar{\theta}_{t-1}})(1 - \mathrm{e}^{-2\theta_t'})}{1 - \mathrm{e}^{-2\bar{\theta}_t}}. \quad (7)$$

Note that the posterior mean $\tilde{\mu}_t(x_t, x_0)$ has exactly the same form as the optimal reverse path x_{t-1}^* in Eq. (4), meaning that sampling from this posterior distribution is also optimal for recovering the initial state, i.e. the HQ image.

In addition, combining the reparameterization trick ($x_t = m_t + \sqrt{v_t}\,\epsilon_t$) with the noise prediction network $\tilde{\epsilon}_\phi(x_t, \mu, t)$ gives us a simple way to estimate x_0 at time t:

$$\hat{x}_0 = \mathrm{e}^{\bar{\theta}_t} \big(x_t - \mu - \sqrt{v_t}\tilde{\epsilon}_\phi(x_t, \mu, t)\big) + \mu, \quad (8)$$

where m_t and v_t are the forward mean and variance in Eq. (2), and ϕ is the learnable parameters. Then we iteratively sample reverse states based on this posterior distribution starting from noisy LQ images for efficient restoration.

4. Experiments

We provide evaluations on different image restoration tasks to illustrate the effectiveness of the proposed method.

Implementation Details For all experiments, we use the AdamW [27] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 2×10^{-4} and decayed to 1e-6 by the Cosine scheduler for 500 000 iterations. The noise level is fixed to 50 and the number of diffusion denoising steps is set to 100 for all tasks. We set the batch size to 16 and the training patches to 256×256 pixels. All models are implemented with PyTorch [38] and trained on a single A100 GPU for about 3-4 days.

4.1. Evaluation of IR in the Wild

Datasets and Metrics We train our model on the LSDIR dataset [21] which contains 84 991 high-quality images with rich textures and their downsampled versions. In training, we only utilize the collected HQ images and synthetically generate all HQ-LQ image pairs following the proposed degradation pipeline in Fig. 2. In testing, we evaluate our model on two external datasets: DIV2K [1] and RealSR $\times 2$ [3]. Specifically, the DIV2K dataset contains 100 2K resolution image pairs with all LQ images generated using our degradation pipeline, while the RealSR $\times 2$ dataset contains 30 high-resolution real-world captured image pairs. In both datasets, we upscale all LQ images to have the same size as the corresponding HQ images for $\times 1$ image restoration. For wild IR, we pay more attention to the visual quality of restored images and thus prefer to compare perceptual metrics such as LPIPS [67], DISTS [8], FID [12], and NIQE [35]. Note that NIQE is a non-reference metric that only evaluates the quality of the output. In addition, we also report distortion metrics like PSNR and SSIM since we also want the prediction to be consistent with the input.

Comparison Approaches We compare our method DACLIP-IR with other state-of-the-art photo-realistic wild image restoration approaches: Real-ESRGAN [53], StableSR [51], SeeSR [56], and SUPIR [57]. All these comparison methods use the same degradation pipeline as that in Real-ESRGAN. Moreover, StableSR, SeeSR, and SUPIR employ pretrained Stable Diffusion models [39, 44] as diffusion priors for better generalization on out-of-distribution images. SeeSR and SUPIR further leverage powerful vision-language models (RAM [70] and LLaVA [26], respectively) to provide additional textual prompt guidance for image restoration in the wild.

Results The quantitative results on the DIV2K and RealSR $\times 2$ datasets are summarized in Table 1 and Table 2, respectively. It is observed that DACLIP-IR achieves the best performance over all approaches on the two datasets. The results are quite expected for the DIV2k dataset since we use the same degradation pipeline in both training and testing. For the RealSR $\times 2$ images, their degradations are unseen for all approaches and our DACLIP-IR still outperforms other methods on most metrics. Moreover, one can observe that changing the degradation pipeline directly decreases the performance on both datasets. And it is worth noting our SDE model is trained from scratch while all other diffusion-based approaches (StableSR, SeeSR, and SUPIR) leverage pretrained Stable Diffusion models as priors, demonstrating the effectiveness of the proposed method and our new degradation pipeline.

A visual comparison of the proposed method with other

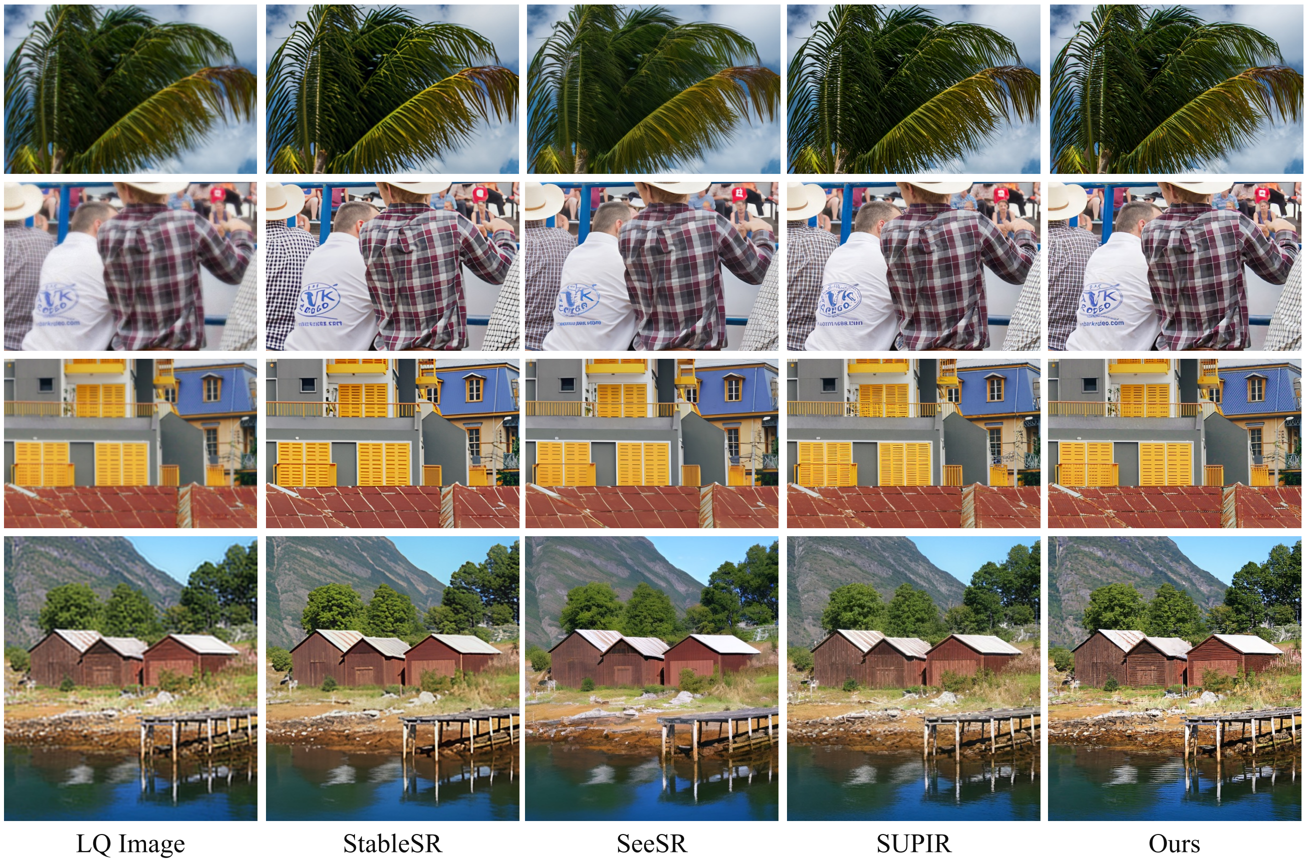

Figure 5. Visual comparison of the proposed model with other state-of-the-art photo-realistic image restoration approaches on our synthetic DIV2K [1] dataset. Our method trains the diffusion model from scratch while other approaches leverage pretrained Stable Diffusion models. Note that all methods using Stable Diffusion are prone to generate unrecognizable text, such as for the white shirt in the second row.

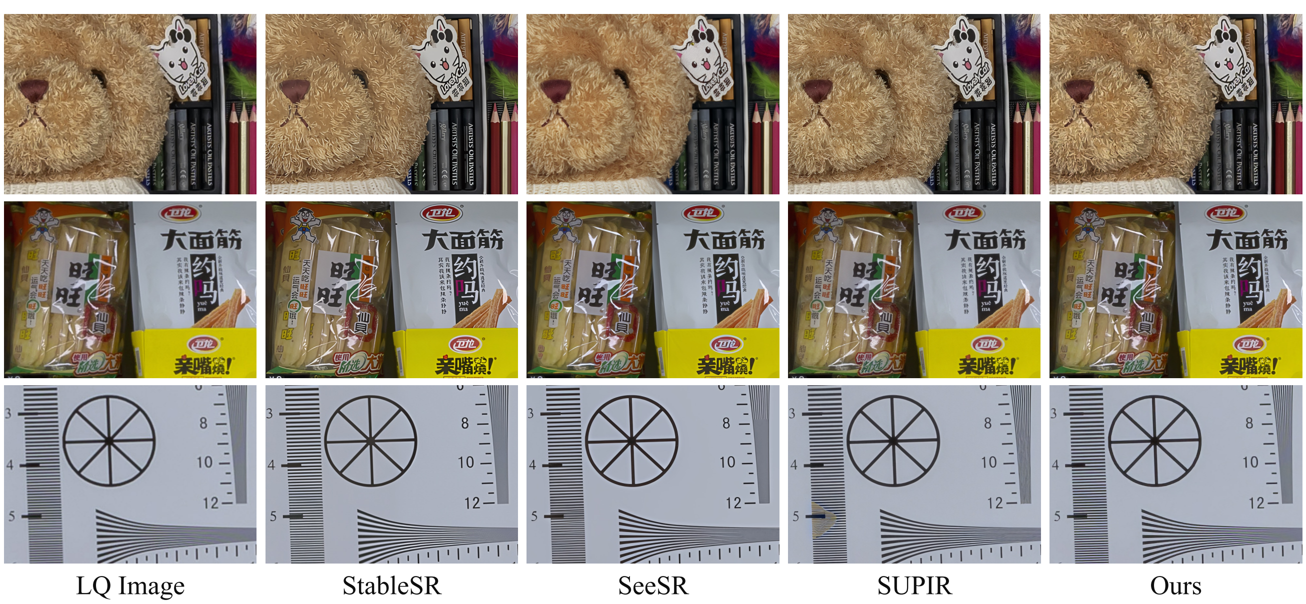

Figure 6. Visual comparison of the proposed model with other state-of-the-art photo-realistic image restoration approaches on the RealSR $\times 2$ [3] dataset. Our method trains the diffusion model from scratch while other approaches leverage pretrained Stable Diffusion models.

Table 1. Quantitative comparison between the proposed method with other real-world image restoration approaches on our synthetic DIV2K [1] test set. '†' means that our model is trained with the Real-ESRGAN [53] degradation pipeline.

Method	Distortion		Perceptual			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓
Real-ESRGAN [53]	27.71	0.810	0.200	0.107	27.32	4.41
StableSR [51]	26.04	0.759	0.241	0.123	34.74	4.11
SeeSR [56]	26.29	0.721	0.223	0.114	27.94	3.56
SUPIR [57]	26.81	0.741	0.194	0.099	21.73	3.52
DACLIP-IR†	27.56	0.796	0.195	0.113	24.32	3.43
DACLIP-IR (Ours)	**29.93**	**0.837**	**0.153**	**0.085**	**15.94**	**3.24**

Table 2. Quantitative comparison between the proposed method with other real-world IR approaches on the RealSR $\times 2$ [3] test set. All inputs are pre-upsampled with scale factor 2. '†' means our model trained with the Real-ESRGAN [53] degradation pipeline.

Method	Distortion		Perceptual			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓
Real-ESRGAN [53]	28.03	0.855	0.151	0.117	47.65	4.84
StableSR [51]	27.55	0.838	0.169	**0.112**	54.87	5.45
SeeSR [56]	28.38	0.815	0.212	0.139	40.85	4.20
SUPIR [57]	29.32	0.826	0.175	0.122	31.75	4.61
DACLIP-IR†	28.92	0.858	0.184	0.138	33.76	**4.19**
DACLIP-IR (Ours)	**30.65**	**0.878**	**0.148**	0.113	**30.09**	4.31

state-of-the-art photo-realistic IR approaches on the two datasets is illustrated in Fig. 5 and Fig. 6. One can see that all these methods can restore visually high-quality images. Moreover, results produced by SeeSR and SUPIR seem to have more details than StableSR, indicating the importance of textual guidance in diffusion-based image restoration. But in terms of distortion metrics that measure the consistency w.r.t the input, we found that pretrained Stable Diffusion models might introduce unclear priors and thus tend to generate text stroke adhesion which is unrecognizable, for example on the back of the white shirt in the second-row of Fig. 5. And in some cases, the SUPIR further produces fake textures and block artifacts, as shown in the third-row of Fig. 5 (the yellow window frames) and the third-row of Fig. 6 (the weird block around '5'). Although our method trains the diffusion model from scratch, its results still look realistic and are consistent with the inputs.

Results on the NTIRE Challenge We also evaluate our model on the NTIRE 2024 'Restore Any Image Model (RAIM) in the Wild' challenge [24], as shown in Table 3. To generalize to the challenge dataset, we first train our model on LSDIR [21] with synthetic image pairs, and then finetune it on a mixed dataset that contains both synthetic and real-world images from LSDIR [21] and RealSR [3]. Note that we use the same model for both phase two and phase three of the challenge, but employ the original reverse-time SDE sampling in phase three for better visual results (small

Table 3. Final results of the NTIRE 2024 RAIM challenge.

Team	Phase 2	Phase 3	Final Score	Rank
MiAlgo	79.13	57	91.65	1
Xhs-IAG	81.96	47	82.07	2
So Elegant	79.69	46	80.09	3
IIP IR	80.03	14	45.94	4
DACLIP-IR	78.65	9	40.03	5
TongJi-IPOE	72.99	11	39.91	6
ImagePhoneix	78.93	4	34.79	7
HIT-IIL	69.80	1	27.92	8

Figure 7. Inpainting results on a web-downloaded face image.

noise makes the photo look more realistic).

4.2. Effectiveness of the Posterior Sampling

This section adopts the same settings as the DACLIP [32] and focuses on unified image restoration (UIR) which trains and evaluates a single model on multiple IR tasks.

Robust DACLIP Model Notice that the original DACLIP is sensitive to input degradations since it is trained on specific datasets without data augmentation. To address this issue, we follow the synthetic training idea from wild image restoration and propose a robust DACLIP model. Similar to the original DACLIP, this robust model is trained on 10 datasets for unified image restoration. However, we now also add mild degradations such as noise, resize, and JPEG compression (first part of the degradation pipeline in Fig. 2) to the LQ images for data augmentation. The resulting model can then better handle real-world inputs that contain minor corruptions. Fig. 7 shows a face inpainting comparison for a web-downloaded image example. As one can see, the original DACLIP model completely fails to inpaint this image. On the other hand, the robust DACLIP model restores the face well, illustrating its robustness.

Evaluation and Analysis To analyze the effectiveness of the proposed posterior sampling, we choose 3 (out of 10) tasks for evaluation: raindrop removal on the RainDrop [41] dataset, low-light enhancement on the LOL [54] dataset, and color image denoising on the CBSD68 [33] dataset. The comparison methods include recent all-in-one image restoration approaches: AirNet [17], PromptIR [40], IR-SDE [30], and the original DACLIP [32]. Finally, the posterior sampling is applied to our robust DACLIP model. The comparison results are reported in Table 4. The PromptIR performs better on distortion metrics

Table 4. Comparison of different methods on the unified image restoration task. 'robust' means we add mild synthetic degradations (e.g., resize, noise, and JPEG) to LQ images in training as a data augmentation strategy for out-of-distribution data generalization. '*' means the method uses the proposed optimal posterior sampling approach for image generation. Here we report the results on the RainDrop [41], LOL [54], and CBSD68 [33] datasets for raindrop removal, low-light enhancement, and denoising task evaluation, respectively.

Method	RainDrop [41]				LOL [54]				CBSD68 [33]			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
AirNet [17]	30.68	0.926	0.095	52.71	14.24	0.781	0.321	154.2	27.51	0.769	0.264	93.89
PromptIR [40]	31.35	**0.931**	0.078	44.48	**23.14**	0.829	0.140	67.15	**27.56**	**0.774**	0.230	84.51
IR-SDE [30]	28.49	0.822	0.113	50.22	16.07	0.719	0.185	66.42	24.82	0.640	0.232	79.38
DACLIP [32]	30.81	0.882	0.068	38.91	22.09	0.796	0.114	52.23	24.36	0.579	0.272	64.71
DACLIP-robust	30.82	0.869	0.078	27.96	22.05	0.782	0.136	51.01	23.90	0.543	0.310	74.83
DACLIP-robust*	**31.68**	0.921	**0.051**	**21.92**	22.78	**0.848**	**0.092**	**41.50**	25.86	0.723	**0.167**	**62.12**

Figure 8. Visual comparison of the proposed posterior sampling for the DACLIP model on the unified IR task.

(PSNR and SSIM) while diffusion-based approaches have better perceptual performance (LPIPS and FID). Although the robust DACLIP model involves more degradations in training, it still performs similarly to its original version. By using the proposed posterior sampling in inference, the performance of the robust DACLIP model is significantly improved across all metrics and tasks. Especially for the denoising task, posterior sampling leads to the best LPIPS and FID performance, proving its effectiveness.

5. Conclusion

This paper addresses the problem of photo-realistic image restoration in the wild. Specifically, we present a new degradation pipeline to generate low-quality images for synthetic data training. This pipeline includes diverse degradations (e.g., different blur kernels) and a random shuffle strategy to increase the generalization. Moreover, we improve the degradation-aware CLIP by adding multiple degradations to the same image and minimizing the embedding distance between LQ-HQ image pairs to enhance the LQ image embedding. Subsequently, we present a posterior sampling approach for IR-SDE, which significantly improves the performance of unified image restoration. Finally, we evaluate our model on various tasks and the NTIRE RAIM challenge and the results demonstrate that the proposed method is effective for image restoration in the wild.

Acknowledgements This research was partially supported by the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by the Knut and Alice Wallenberg Foundation, by the project *Deep Probabilistic Regression – New Models and Learning Algorithms* (contract number: 2021-04301) funded by the Swedish Research Council, and by the *Kjell & Märta Beijer Foundation*. The computations were enabled by the *Berzelius* resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 5, 6, 7

[2] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 3

[3] 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*, pages 3086–3095, 2019. 5, 6, 7

[4] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1329–1338, 2022. 2

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 1

[6] Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *Advances in Neural Information Processing Systems*, 36, 2023. 1

[7] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 2, 5

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 2

[10] 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. 2

[11] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 1

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2

[14] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 1, 2

[15] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 4

[16] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2

[17] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17452–17462, 2022. 2, 7, 8

[18] 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. 1, 2

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[20] Youwei Li, Haibin Huang, Lanpeng Jia, Haoqiang Fan, and Shuaicheng Liu. D2c-sr: A divergence to convergence approach for real-world image super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 379–394. Springer, 2022. 2

[21] 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*, pages 1775–1787, 2023. 5, 7

[22] Wenyi Lian and Shanglian Peng. Kernel-aware burst blind super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4892–4902, 2023. 2

[23] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1

[24] Jie Liang, Radu Timofte, Qiaosi Yi, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Lei Zhang, et al. NTIRE 2024 restore any image model

(RAIM) in the wild challenge: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2024. 7

[25] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 5

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 1

[29] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17642–17652, 2022. 3

[30] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. In *International Conference on Machine Learning*, pages 23045–23066. PMLR, 2023. 1, 2, 3, 4, 7, 8

[31] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691, 2023. 1, 2

[32] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4, 7, 8

[33] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 18th IEEE International Conference on Computer Vision (ICCV)*, pages 416–423. IEEE, 2001. 7, 8

[34] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[35] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 5

[36] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 4

[37] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[39] 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. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2, 5

[40] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 7, 8

[41] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2482–2491, 2018. 7, 8

[42] 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*, pages 8748–8763. PMLR, 2021. 3

[43] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10721–10733, 2023. 1

[44] 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*, pages 10684–10695, 2022. 1, 2, 3, 4, 5

[45] 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*, pages 234–241. Springer, 2015. 2

[46] 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. 2

[47] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 1

[48] 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, 2022. 1, 2

[49] 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*, 2020. 3

[50] 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. *arXiv preprint arXiv:2311.16512*, 2023. 2

[51] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 1, 2, 5, 7

[52] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1, 2

[53] 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*, pages 1905–1914, 2021. 1, 2, 3, 4, 5, 7

[54] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 7, 8

[55] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 1

[56] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. *arXiv preprint arXiv:2311.16518*, 2023. 2, 5, 7

[57] 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. *arXiv preprint arXiv:2401.13627*, 2024. 1, 2, 5, 7

[58] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum. Image deblurring with blurred/noisy image pairs. *ACM Transactions on Graphics (TOG)*, 26(3):1–es, 2007. 4

[59] Conghan Yue, Zhengwei Peng, Junlong Ma, Shiyan Du, Pengxu Wei, and Dongyu Zhang. Image restoration through generalized ornstein-uhlenbeck bridge. *arXiv preprint arXiv:2312.10299*, 2023. 1

[60] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2023. 1

[61] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. 1, 2

[62] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 1

[63] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017.

[64] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021. 2

[65] 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*, pages 4791–4800, 2021. 2, 3

[66] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[67] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 2, 5

[68] Ruoqi Zhang, Ziwei Luo, Jens Sjölund, Thomas B Schön, and Per Mattsson. Entropy-regularized diffusion policy with q-ensembles for offline reinforcement learning. *arXiv preprint arXiv:2402.04080*, 2024. 4

[69] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2

[70] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 5

[71] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3