Towards Unsupervised Blind Face Restoration using Diffusion Prior

Tianshu Kuai[2,†], Sina Honari[1], Igor Gilitschenski[2,3], Alex Levinshtein[1]
[1]Samsung AI Center Toronto, [2]University of Toronto, [3]Vector Institute for AI

Abstract

Blind face restoration methods have shown remarkable performance, particularly when trained on large-scale synthetic datasets with supervised learning. These datasets are often generated by simulating low-quality face images with a handcrafted image degradation pipeline. The models trained on such synthetic degradations, however, cannot deal with inputs of unseen degradations. In this paper, we address this issue by using only a set of input images, with unknown degradations and without ground truth targets, to fine-tune a restoration model that learns to map them to clean and contextually consistent outputs. We utilize a pre-trained diffusion model as a generative prior through which we generate high quality images from the natural image distribution while maintaining the input image content through consistency constraints. These generated images are then used as pseudo targets to fine-tune a pre-trained restoration model. Unlike many recent approaches that employ diffusion models at test time, we only do so during training and thus maintain an efficient inference-time performance. Extensive experiments show that the proposed approach can consistently improve the perceptual quality of pre-trained blind face restoration models while maintaining great consistency with the input contents. Our best model also achieves the state-of-the-art results on both synthetic and real-world datasets. Project Page.

1. Introduction

Image restoration is a fundamental task in computational photography that aims to recover a high-quality image from its degraded low-quality counterpart. Blind image restoration is a more challenging task, where the degradation process is unknown. One needs to find a good balance between maintaining the fidelity of the image content and the output's perceptual quality. This is particularly important in the case of blind face restoration, as both fidelity and quality are important when restoring face images.

Most of the existing blind face restoration methods [19, 62, 65, 72, 79] are trained in a supervised manner using a paired dataset of low-quality inputs and high-quality target images. The training pairs are often constructed by manually designing a degradation process [31, 33, 62], where a high-quality image is synthetically degraded to form the corresponding low-quality input. Supervised learning achieves great performance on test data that aligns with the training degradations (Fig. 1(a)). However, this produces results with severe artifacts when tested on images that do not fall under the training degradation distribution (Fig. 1(b)). In addition, ground-truth data is not always available for the supervised learning setup. We commonly have access to only the low-quality observations in a real-world setting. In this paper, we address such blind unsupervised setup with neither access to the paired ground truth images nor to the degradation process of the inputs.

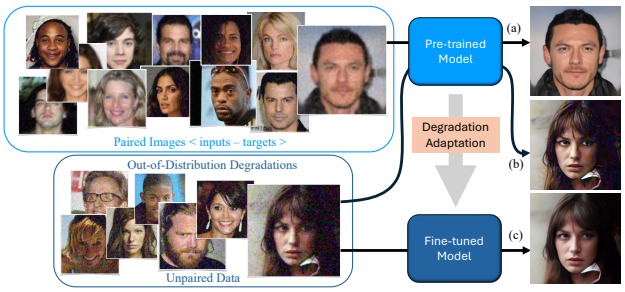

Figure 1. **Overview.** Given a restoration model pre-trained on synthetic datasets in a supervised fashion, it can produce high-quality restoration on low-quality images that are aligned with the degradation distribution used in training (a). However, it often fails on inputs of out-of-distribution degradations (b). We propose an unsupervised pipeline to adapt a pre-trained model to unpaired degraded images of the target degradation with a much smaller data size. This addresses the domain gap in degradation types without paired ground-truth images or the knowledge of the target data's degradation type (c). (**zoom in for details**).

Image diffusion models [22, 49] have recently shown remarkable performance in image generation. Due to their powerful modeling of the natural image manifold, pre-trained diffusion models can be used as priors for image restoration tasks in a zero-shot manner. Some blind image restoration methods [9, 37, 66, 69, 75] have achieved great results on severely degraded data. However, they require

[†]Work done during an internship at Samsung AI Center Toronto.

running the diffusion model's sampling process during inference, which results in a significant computational cost and extremely slow runtime. On the other hand, a series of works [6, 13, 26, 39, 40, 63, 82] involve designing handcrafted denoising process for known degradation types in a supervised setup, which limits the applicable scenarios.

In this work, we tackle the problem of unsupervised blind image restoration by taking the advantages from the above two groups of works. Given a pre-trained restoration model that fails on inputs with some out-of-distribution degradations (target degradations), our approach consists of two stages: pseudo target generation and model fine-tuning. To generate pseudo targets, we design a denoising diffusion process that cleans up the restoration model's outputs, where it preserves the input image content while considerably enhancing high frequency details. In the second step, the cleaned images are treated as pseudo targets to fine-tune the pre-trained restoration model using input and pseudo target pairs. The fine-tuned model is then able to handle inputs with the same target degradations (Fig. 1(c)). The proposed approach requires only a small set of unpaired low-quality observations for training, and does not require running the diffusion model at test time, which is a much more practical setting for real-world applications. To the best of our knowledge, this is the first approach to use pseudo-targets to adapt a pre-trained restoration model to unknown degradations for blind face restoration.

In summary, our contributions are as follows: (i) An unsupervised pipeline for adaptation of face restoration models to unknown unpaired degradations; (ii) a method to obtain content-preserving pseudo targets from a diffusion model that achieves better fidelity and perceptual quality than previous zero-shot diffusion-based restoration methods; (iii) our approach consistently improves the pre-trained models, and our best fine-tuned model achieves state-of-the-art performance on both synthetic and real-world datasets without the need for running a diffusion model at inference time.

2. Related Work

2.1. Supervised Blind Face Restoration

Most of the existing methods in blind face restoration involve supervised learning with simulated training data pairs, combined with various types of priors. Due to the structured nature of facial images, many works explore face-specific priors such as geometry priors [2, 3, 30, 57, 73, 74, 80, 81] and reference priors [10, 32–34] to retain natural and faithful restoration for the given low-quality faces. To further improve the perceptual quality of the restoration, several models [1, 35, 44, 62, 72] employ GAN-based priors with perceptual and adversarial losses during supervised training, or use pre-trained GAN models [24, 25] as priors directly. Some works [31, 34] explore facial component dictionaries as a more robust prior for higher quality restoration and identity preservation.

Following the high-quality codebook learning approaches [12, 61], more recent methods [19, 65, 78, 79] show great performance on real-world degraded faces by first pre-training on large-scale clean face data to obtain high-quality discrete dictionaries or codebooks as priors during restoration. These methods achieve great perceptual quality, but do not generalize to the degradations that are outside of the training degradation distribution. The reliance on the large-scale training data pairs also limit their practicality for real-world applications. Different from the previous supervised approaches, we aim to address the generalization problem of pre-trained face restoration models on out-of-distribution degradations in an unsupervised manner.

2.2. Diffusion Priors for Image Restoration

Diffusion models [22, 49] have recently become the most powerful generative models in image synthesis. Despite not being initially designed for low-level imaging tasks, some works modify diffusion model's architecture and train them by conditioning on either the degraded image or its features for tasks such as super-resolution (SR) [17, 41, 52, 54, 68], shadow-removal [20, 41], deblurring [48, 68], inpainting and uncropping [53, 68], face-restoration [77], and adverse weather restoration [41, 46]. Some works [40, 47, 67] explore different sampling and training procedure of diffusion models for better restoration performance. However, all of these models require supervised training with large amount of data pairs and computational resources, while still suffering from lack of adaptability to generalize to out-of-distribution degradations.

A large group of the methods utilize pre-trained diffusion models for zero-shot image restoration tasks [5–7, 11, 13, 14, 26–28, 45, 51, 60, 63, 82] including super-resolution, inpainting, deblurring, denoising, and JPEG artifact correction. However, they require the knowledge of the degradation type to design custom denoising process, which cannot be applied to blind image restoration directly. To tackle the blind restoration problem, where the degradation is unknown, the conditioning or guidance during the denoising process needs to be generalized enough to handle a variety of degradation types. On this line of research, some papers [4, 16, 66] use the low frequency content of the images to guide the denoising process in order to preserve the content of the image. Some other papers [37, 69, 75] use a simple pre-trained restoration model to first reduce the amount of artifacts in the image while preserving the smoothed semantics, and then use a pre-trained or fine-tuned diffusion model to inject sharp textures to the restoration model's output. Although these zero-shot approaches achieve great level of perceptual quality, they share the common problem

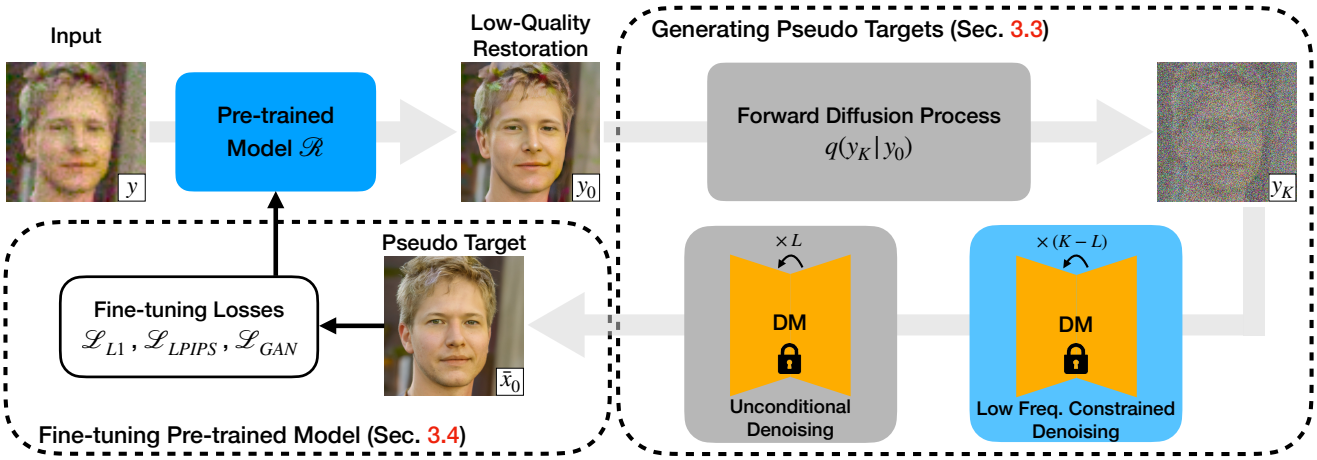

Figure 2. **Overview of our unsupervised fine-tuning pipeline.** Given a pre-trained restoration model that produces low-quality restoration outputs (severe artifacts on hair and over-smoothed skin) on samples with unknown and out-of-distribution degradations, we generate pseudo targets using a pre-trained unconditional diffusion model with a combination of low frequency content constrained denoising and unconditional denoising. The generated clean images can be used as pseudo GT to fine-tune the pre-trained restoration model without the need for real GT images.

of long inference time due to running the diffusion model for every input. There have been efforts to make the diffusion models faster by reducing the number of sampling using DDIM [59] or progressively distilling it into fewer steps [55] at the expense of image quality. Our approach elimiates the burden of running the diffusion model altogether at inference time. It only uses the diffusion model's outputs during training as pseudo targets to improve a pre-trained restoration model by injecting its prior to restore unknown degradations.

3. Method

3.1. Preliminaries on Diffusion Models

Denoising Diffusion Probabilistic Model (DDPM) [22, 58] has been one of the most used and powerful generative models in computer vision. An unconditional diffusion model learns a natural image manifold from large-scale image datasets. It follows a Markov forward process to gradually corrupt an image x_0 with a pre-defined Gaussian noise variance schedule β_t for each timestep $t \in \{1, 2, ..., T\}$. Thanks to the Markovian formulation of the diffusion process, one can write the expression for the noisy image x_t at any timestep t, given the clean image x_0 as

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

An unconditional diffusion model generates natural images by reversing the forward diffusion process. Specifically, the process can be written as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \boldsymbol{\mu}_\theta(x_t, t), \sigma_t^2\mathbf{I}), \tag{2}$$

where σ_t^2 is set to a time-dependent constant, and the mean of the denoised image $\boldsymbol{\mu}_\theta(x_t, t)$ is computed as:

$$\boldsymbol{\mu}_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(x_t, t)\Big), \tag{3}$$

where the noise $\boldsymbol{\epsilon}_\theta(x_t, t)$ at timestep t is predicted by a trained timestep conditioned U-Net [50]. To perform unconditional image generation, we can start with a sample from the standard Gaussian distribution $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and gradually denoise it using the predicted noise at each timestep. Note that one can use techniques from DDIM [59] or simply uniformly skip timesteps during the reverse diffusion process to accelerate the denoising process.

3.2. Method Overview

Given a restoration model that is pre-trained on synthetic data pairs, its performance on out-of-distribution inputs is often heavily degraded. To bridge the gap without the need for paired ground-truth high-quality images, we use a pre-trained unconditional diffusion model [22] to clean up the artifacts in restoration model's output via a low frequency constrained denoising process. As shown in Fig. 2, our pipeline consists of two stages: 1) generating pseudo targets using a diffusion model (Section 3.3) and 2) using the generated targets to fine-tune the pre-trained restoration model (Section 3.4).

3.3. Generating Pseudo Targets

Consider a pre-trained restoration model, $\mathcal{R}$, and a real-world low-quality image observation y. Due to the domain gap between the synthetic data and the real-world data, the output from a pre-trained restoration model, $y_0 = \mathcal{R}(y)$,

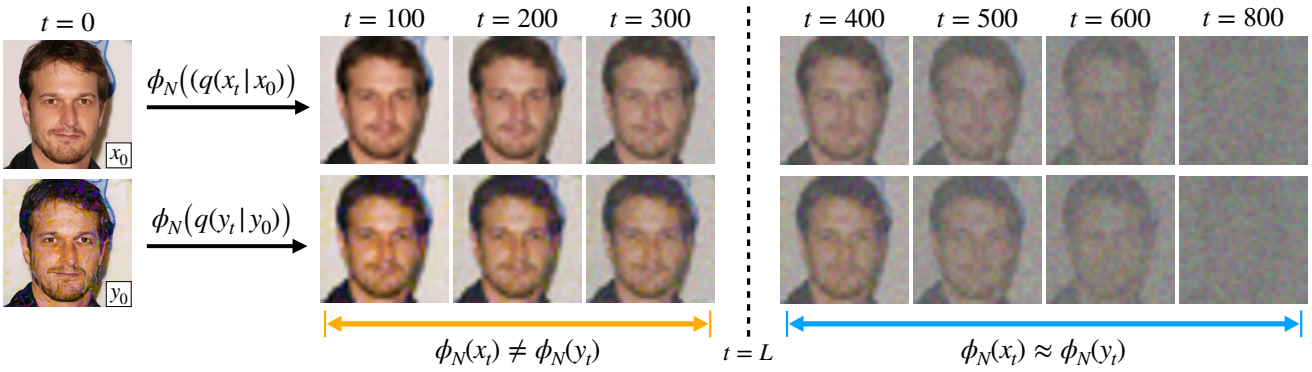

Figure 3. **Visualization of low frequency contents at different timesteps.** We show low frequency contents of the low-quality restoration from a pre-trained SwinIR [36] and its GT counterparts at different timesteps of the forward diffusion process (**zoom in for details**).

still contains a lot of artifacts. Following Eq. (1), we can apply the forward diffusion process on y_0, to get a noisy version of the low-quality restoration y_t at timestep t as:

$$y_t = \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (4)$$

If we directly perform an unconditional denoising on the noisy image y_t as in [75], the structured content will not be preserved well, which yields inconsistent restoration. Similar to the observations from [9, 66], we found that as more Gaussian noise (larger timestep) is added to the low-quality restoration, the low frequency content of y_t is getting closer to the low frequency content of the noisy image x_t as if we start with the clean image counterpart x_0. Specifically, given a low pass filter ϕ_N and if t is large enough ($t > L$):

$$\phi_N(y_t) \approx \phi_N(x_t) = \phi_N(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}), \quad (5)$$

where x_t is the noisy version of the clean image, and $\boldsymbol{\epsilon}$ is the same sampled noise in Eq. (4). We visualize and compare the low frequency contents of the two images at different timesteps in Fig. 3. The noisy images become closer as timestep increases, and eventually indistinguishable visually after $t = 400$. With this critical observation, we can constrain the denoising process by regularizing the low frequency content at each denoising timestep when $t > L$, in order to preserve the structural information. At lower timesteps ($t <= L$), the low frequency property in Eq. 5 no longer holds and applying such regularization will deteriorate the denoising process. In addition, since we are using an unconditional diffusion model, going all the way to $t = T$ would completely destroy all the information in the image. Hence, we start the low frequency constrained denoising process at a smaller timestep $t = K$, where the low frequency content is not yet destroyed by the injected Gaussian noise.

Combined with the insights above, we describe our pseudo target generation process as follows: we take the restoration model's output, y_0, on an image from the target dataset, and follow the pre-defined noise schedule to inject Gaussian noise into the image up to timestep K. We then pass it to the diffusion model to clean up the image. Similar to [4, 16, 66], we guide the denoising process by constraining the low frequency content to be consistent with the input. This is done by replacing the low frequency content of the denoised image with the corresponding part from the noisy copy of the input image at each time step. Some methods [4, 16] apply such guidance on all denoising steps, which would lead to blurry outputs with artifacts due to over-constraining the denoised images on information that can be a mixture of signal and noise.

Different from previous methods, we only apply this low frequency content constraint for timesteps when $t > L$. This is because the low frequency property does not hold anymore for small timesteps ($t \leq L$) as shown in Fig. 3. Moreover, we observe that the denoised images at these timesteps already have reasonably good structure. There-

Algorithm 1 Generating pseudo targets using a pre-trained diffusion model

Input: low-quality restoration output $y_0 = \mathcal{R}(y)$, low-pass filter ϕ_N
Output: pseudo target $\bar{x}_0$ for low-quality input y
$\bar{x}_K \leftarrow$ sample from $\mathcal{N}(y_K; \sqrt{\bar{\alpha}_K} y_0, (1-\bar{\alpha}_K)\mathbf{I})$
for t from K to 1 **do**
  $\bar{x}_{t-1} \leftarrow$ sample from $p_\theta(\bar{x}_{t-1}|\bar{x}_t)$ ▷ unconditional denoising
  if $t > L$ **then**
    $y_{t-1} \leftarrow$ sample from $\mathcal{N}(y_{t-1}; \sqrt{\bar{\alpha}_{t-1}} y_0, (1-\bar{\alpha}_{t-1})\mathbf{I})$
    $\bar{x}_{t-1} \leftarrow \bar{x}_{t-1} - \phi_N(\bar{x}_{t-1}) + \phi_N(y_{t-1})$ ▷ low frequency content constraint
  end if
end for
return $\bar{x}_0$

fore, we perform unconditional denoising for the remaining L timesteps since unconditional denoising steps contribute to high-frequency details at small timesteps [43]. With this approach, there is no need for directly estimating x_0 from x_L in one step and running another enhancement model on the generated image [66]. We summarize the detailed procedure of generating pseudo targets in Algorithm 1.

3.4. Fine-tuning the Pre-trained Models

After obtaining the pseudo targets from the diffusion model, one can fine-tune the pre-trained restoration model with the low-quality inputs and pseudo targets data pairs. We apply the widely used image-level $L1$ loss, perceptual (LPIPS) loss [76], and adversarial (GAN) loss [18]:

$$\mathcal{L}_{L1} = ||\mathcal{R}(y) - \bar{x}_0||_1, \quad \mathcal{L}_{LPIPS} = \text{LPIPS}\big(\mathcal{R}(y), \bar{x}_0\big), \\ \mathcal{L}_{GAN} = \log\big(1 - \mathcal{D}(\mathcal{R}(y))\big), \tag{6}$$

where $\mathcal{R}$ is our restoration model, and $\mathcal{D}$ is a discriminator that outputs the probability of its input coming from the distribution of real natural faces. This discriminator is optimized from scratch along with the restoration model, with the following cross-entropy training objective [18]:

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{x\sim\mathcal{R}(y)}\big[-(1 - \log\mathcal{D}(x))\big] + \mathbb{E}_{x\sim\mathbb{P}_r}\big[-\log\mathcal{D}(x)\big], \tag{7}$$

where $\mathbb{P}_r$ represents the distribution of real high-quality face images. In practice we treat the images in FFHQ dataset [24] as our real data distribution. Therefore, we use randomly sampled images from the FFHQ dataset as clean references for optimizing the discriminator. This ensures that the discriminator is robust enough to provide useful gradient signals for optimizing the restoration model. The complete training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{L1} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{GAN}\mathcal{L}_{GAN}, \tag{8}$$

where λ_{LPIPS} and λ_{GAN} are hyperparameters for the weights of the losses.

4. Experiments

4.1. Implementation and Evaluation Settings

Pre-trained Models. We use a pre-trained unconditional face diffusion model with image resolution of 512×512. This diffusion model is trained by following the training procedure from [8,75] on the entire FFHQ dataset [24] with 70,000 images. We demonstrate the effectiveness of our approach on two model architectures: SwinIR [36] and CodeFormer [79] (current state-of-the-art non-diffusion based method for blind face restoration). For SwinIR, we follow training setup from [62] to pre-train the SwinIR model on the FFHQ [24] dataset following the losses in Eq. (8). For CodeFormer, we use the pre-trained checkpoint from the authors. Both models are trained using the synthetic degradation function as in [31, 33, 62]:

$$\mathcal{I}_{LQ} = \Big\{\big[(\mathcal{I}_{HQ} * \mathbf{k}_\sigma)_{\downarrow_r} + \mathbf{n}_\delta\big]_{JPEG_q}\Big\}_{\uparrow_r}, \tag{9}$$

where the high-quality image $\mathcal{I}_{HQ}$ is first convolved with a Gaussian blur kernel $\mathbf{k}_\sigma$ of kernel size σ and downsampled by factor of r. Then Gaussian noise of standard deviation δ is added, followed by JPEG compression of quality factor q and upsampling by a factor of r to obtain the low-quality image I_{LQ}.

Low Pass Filter and Timesteps. We adopt the low pass filter from [4], where it is implemented as a sequence of downsampling and upsampling with factor of N. For the SwinIR pseudo targets, we set N to be 16. For the CodeFormer pseudo targets, we set N to be 8 for all the 4× data setup, and to be 16 for all the 8×. We set the starting timestep to be $K = 600$ and apply the low frequency constrained denoising process for 240 timesteps. As a result, we start the unconditional denoising at the timestep of $L = 360$.

Fine-tuning Setup. For fine-tuning the SwinIR model, we set the weights of the losses to be $\lambda_{LPIPS} = 0.1$ and $\lambda_{GAN} = 0.1$ for all the experiments. For CodeFormer [79], we follow their training setup and empirically found that only adopting their code-level losses to optimize the code prediction module and the VQ-GAN encoder gives better fine-tuning performance than the image-level losses in Eq. (8). Note that this fine-tuning procedure is specific to CodeFormer architecture and cannot be generalized to all model architectures. Please refer to the supplementary material for more details on our fine-tuning experiments.

Testing Datasets. We evaluate our pipeline on both synthetic and real-world datasets. For the synthetic dataset, we generate low-quality testing inputs using 3,000 high-quality face images from the CelebA-HQ [23] dataset. To simulate more realistic degradations instead of the the commonly used pipeline (Eq. (9)) in [31, 33, 62], we apply the following degradation model:

$$\mathcal{I}_{LQ} = \Big\{ISP\big[ISP^{-1}((\mathcal{I}_{HQ})_{\downarrow_r}) + \mathbf{n}_c\big]\Big\}_{\uparrow_r}, \tag{10}$$

where we first unprocess (ISP^{-1}) the downsampled high-quality images $\mathcal{I}_{HQ}$ to RAW [56], and then apply the widely used camera noise models [15,38,42] $\mathbf{n}_c$ to simulate noisy RAW images. We then render (ISP) the images back to sRGB [56] and upsample the images to the same resolution as the high-quality ones. We construct datasets at **4**× and **8**× downsampling factors $\downarrow_r$. For each downsampling factor, we generate degraded images at *moderate* and *severe* noise levels. This yield four sets with each set containing 3,000 input-GT pairs. We use 2,500 images to generate the pseudo targets for fine-tuning, and use the remaining 500 pairs for evaluation. Note that during fine-tuning, we do not

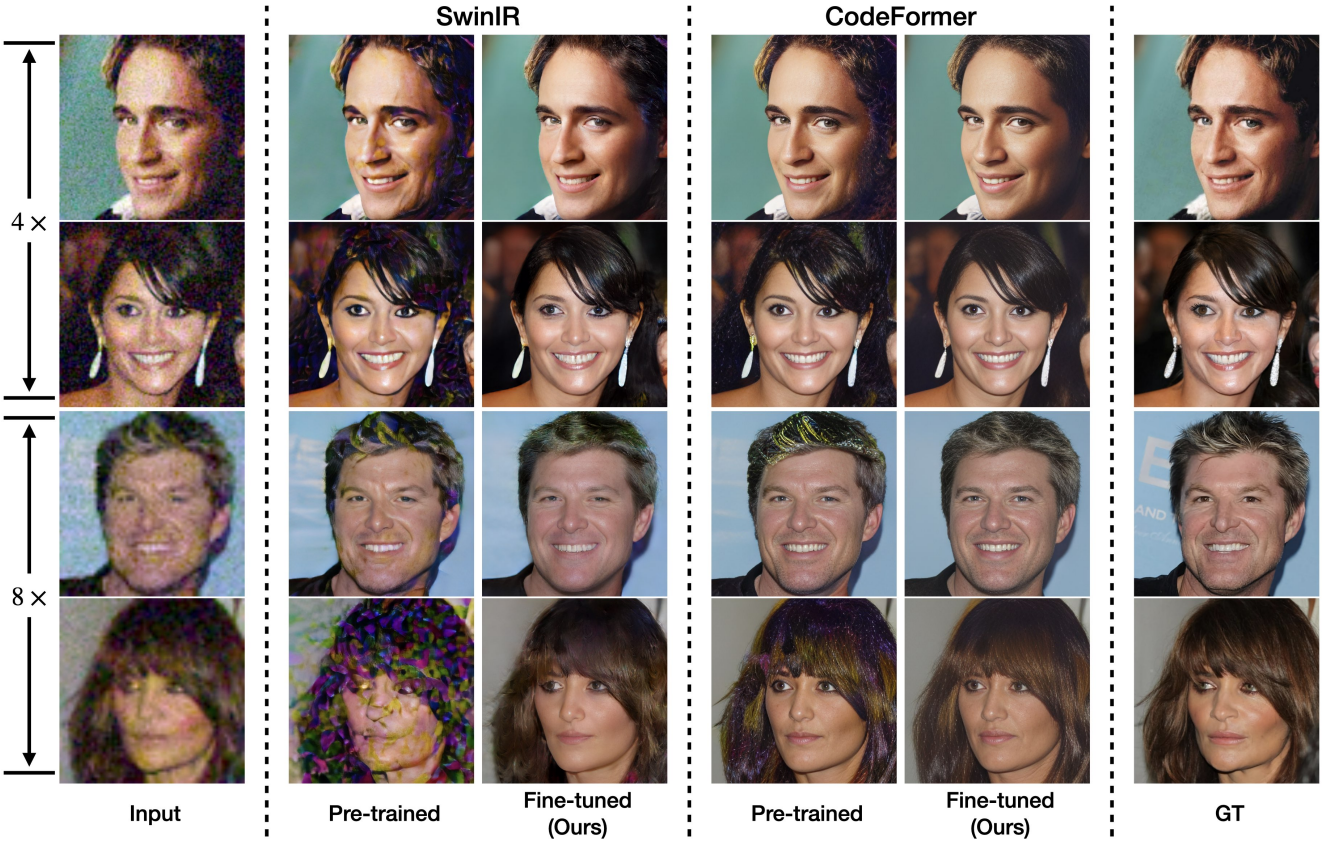

Figure 4. **Qualitative comparison of pre-trained versus fine-tuned models.** We show the effectiveness of our proposed approach to a pre-trained SwinIR [36] and a pre-trained CodeFormer [79] models on 4× and 8× downsampled data at *moderate* noise level. The fine-tuned models are able to produce realistic restoration (**zoom in for details**).

	4× Downsampling				8× Downsampling			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
SwinIR [36]	21.28 / 20.92	0.5744 / 0.5444	0.5446 / 0.5842	74.12 / 120.44	21.28 / 19.00	0.5744 / 0.4813	0.5446 / 0.6435	99.44 / 152.90
SwinIR + Ours	**24.75 / 23.41**	**0.6676 / 0.6284**	**0.3853 / 0.4156**	**41.42 / 49.46**	**23.28 / 21.91**	**0.6206 / 0.5720**	**0.4348 / 0.4821**	**68.25 / 115.97**
CodeFormer [79]	**24.31 / 22.90**	**0.6335** / 0.5810	**0.4007** / 0.4420	**40.66** / 53.02	22.19 / 20.60	0.5716 / 0.5108	0.4420 / 0.4938	51.91 / 72.16
CodeFormer + Ours	23.20 / 22.85	0.6138 / **0.6036**	0.4117 / **0.4258**	41.74 / **41.21**	**22.28 / 21.38**	**0.5848 / 0.5514**	**0.4290 / 0.4589**	**41.72 / 46.66**

Table 1. Improvements gained on at 4× downsampling and 8× downsampling test sets at *moderate* and *severe* noise levels for pre-trained SwinIR [36] and CodeFormer [79]. Numbers presented as: [*moderate* / *severe*].

use the GT images and each set is fine-tuned independently. Due to limited space, we average the results on the *moderate* and *severe* settings when presenting results for each one of **4**× and **8**× downsampling factors. More details on our dataset synthesis pipeline and the detailed results on each one of the four sets are provided in the supplementary material. For the real-world dataset, we use the Wider-Test set from [79] for fine-tuning the pre-trained model, and we select another 200 severely degraded images from the testing set of the Wider-Face dataset [70] as **Wider-Test-200** for evaluating fine-tuned models. Note that our Wider-Test-200 has no overlap with the set we obtained from [79].

Evaluation Metrics. For synthetic datasets, we report PSNR, SSIM [64], and LPIPS [76], as we have access to the ground-truth images. For our Wider-Test-200 set, we report the non-reference image quality metrics (MANIQA [71] and MUSIQ [29]). In addition, we report the commonly used FID scores [21] for all datasets, where we use the distribution of the ground-truth images as the reference statistics for synthetic dataset. Since the Wider-Test-200 does not contain the ground-truth images, we use statistics of the FFHQ dataset [24] as reference to measure the FID score.

4.2. Results

Pre-trained vs. fine-tuned models. We first compare the performance of the pre-trained models with the fine-tuned

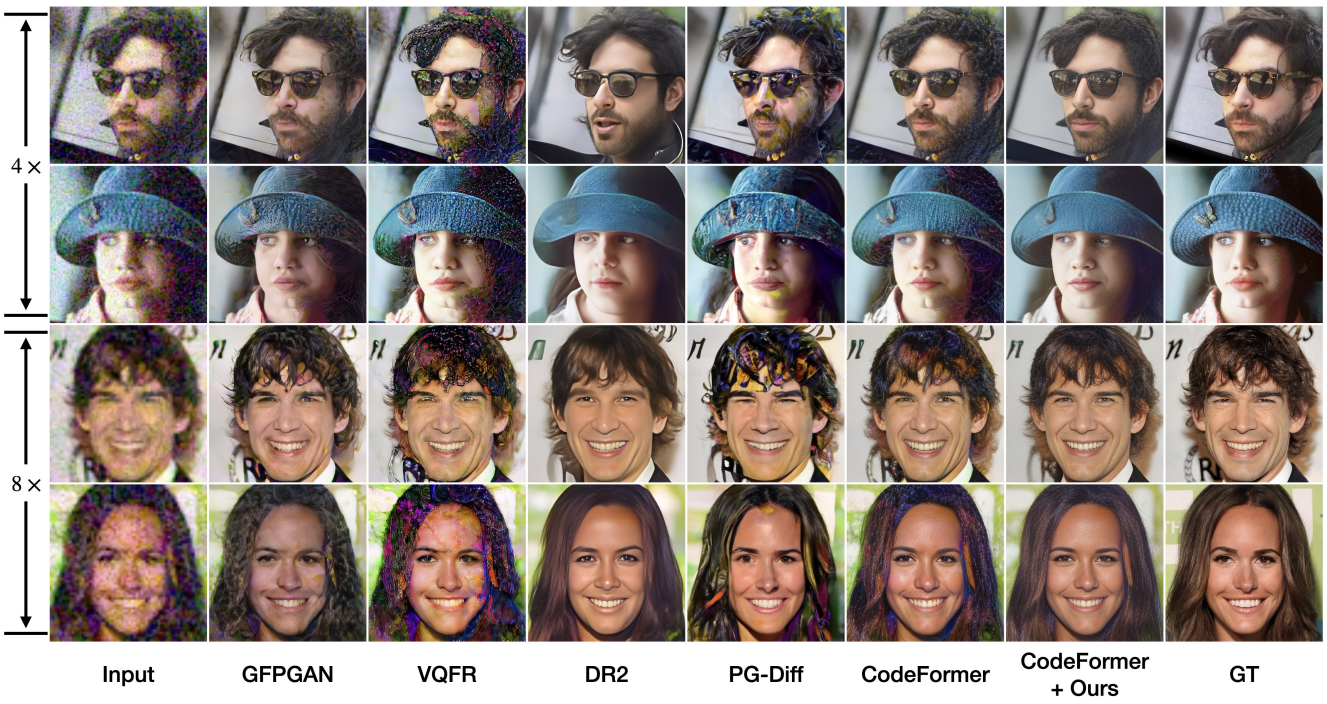

Figure 5. **Qualitative comparison with SOTA baselines on synthetic datasets.** Our fine-tuned CodeFormer model outperforms all other baselines and its pre-trained counterparts on severely degraded inputs from both 4× downsampling and 8× downsampling inputs (**zoom in for details**).

	PSNR↑	SSIM↑	LPIPS↓	FID↓	Diffusion at inference time?
DifFace [75]	20.02	0.5225	0.6077	110.36	✓
DiffBIR [37]	20.28	0.4959	0.6605	126.44	✓
PG-Diff [69]	19.50	0.5339	0.5484	119.82	✓
DR2 [66]	20.34	0.5658	0.4815	54.94	✓
Our pseudo targets	**21.66**	**0.6044**	**0.4706**	**46.29**	✓
GFPGAN [62]	21.09	0.5283	0.5298	82.90	✗
VQFR [19]	19.38	0.4540	0.5567	123.12	✗
CodeFormer [79]	21.40	0.5412	0.4679	62.04	✗
CodeFormer + Ours	**21.83**	**0.5681**	**0.4440**	**44.19**	✗

Table 2. **Results on data at 8× downsampling test set.** Top rows: diffusion-dependent models at test time. Bottom rows: diffusion-free models at test time.

	MANIQA↑	MUSIQ↑	FID↓	Diffusion at inference time?
DifFace [75]	0.5252	55.10	89.19	✓
DiffBIR [37]	0.5994	62.42	92.33	✓
PG-Diff [69]	0.5643	61.69	97.82	✓
DR2 [66]	**0.6007**	**68.05**	90.45	✓
Our pseudo targets	0.5772	62.56	**80.60**	✓
GFPGAN [62]	0.5864	67.14	87.71	✗
VQFR [19]	0.5929	69.19	91.76	✗
CodeFormer [79]	0.6082	66.46	87.60	✗
CodeFormer + Ours	**0.6343**	**73.02**	**84.65**	✗

Table 3. **Results on Wider-Test-200 set.** Top rows: diffusion-dependent models at test time. Bottom rows: diffusion-free models at test time.

ones (obtained following our proposed approach in Section 3). We show the qualitative comparison in Fig. 4.

For SwinIR [36], the pre-trained model produces lots of artifacts. We achieve significant improvements on the perceptual quality of the restoration after fine-tuning. We observe that the pre-trained CodeFormer [79] already outputs faces in relatively good quality at lower degradation levels, but still produces noticeable artifacts on the 8× downsampling data, especially in dark and hair regions. Our fine-tuned model is able to remove such artifacts.

We show quantitative comparison for the two models in Tab. 1. For SwinIR, the pre-trained model receives consistent improvements after fine-tuning. The perceptual improvements are reflected in terms of the large boost in FID and LPIPS scores. For CodeFormer, we found it to be more resistant to smaller degradations (4× *moderate*), and hence the fine-tuning mostly makes the images more realistic by removing the remaining artifacts (as observed in Fig. 4 top two rows). On inputs with larger degradations (4× *severe* and 8×), our approach consistently improves the pre-trained model.

Comparison with state-of-the-art models. We benchmark our best fine-tuned model (fine-tuned CodeFormer) against other blind face restoration baselines in Tab. 2. Our approach (bottom row) outperforms all baseline methods. In top rows we provide results of the zero-shot diffusion-based baselines that use a diffusion model at test time and compare them with our pseudo targets. Our targets improve on all

Number of images	PSNR↑	SSIM↑	LPIPS↓	FID↓	MANIQA↑	MUSIQ↑
Pre-trained	23.57	0.6391	0.4851	74.12	0.4623	64.38
20	22.98	0.6321	0.4607	77.14	0.5239	59.05
100	22.82	0.6218	0.4516	62.60	0.5748	68.56
500	22.84	0.6216	0.4324	53.39	0.5932	72.17
1000	23.10	0.6200	0.4286	51.11	0.5918	72.80
2500	24.75	0.6676	0.3853	41.42	0.6023	73.36

Table 4. **Effects of varying fine-tuning dataset size for SwinIR.**

metrics compared to these models, showing our approach can better preserve the content in the input image (higher PSNR and SSIM) while being more realistic according to LPIPS and FID scores.

The bottom rows in Tab. 2 show test-time non-diffusion based models, compared to which we improve consistently on all metrics. This group contains closer baseline models to ours, as none of them require running a diffusion model at test time, and hence are much more efficient. An interesting observation is that on some metrics our fine-tuned results are better than our pseudo targets. We believe a mixture of inductive bias from the pre-trained model together with distilled information from the diffusion model is injected into the parameters of the restoration model. In addition, directly learning from samples of the target (fine-tuning) dataset helps better generalize to target degradations. As shown in the first ablation study (Sec. 4.3), using more samples of the target dataset leads to better fine-tuned models.

Evaluation on real-world dataset. In Tab. 3, we compare our fine-tuned model with baselines on real-world degradations (Wider-Test-200). We improve on all non-reference based metrics compared to all baselines. We provide qualitative results on Wider-Test-200 in the supplementary material. One observation is that DR2 obtains higher MANIQA and MUSIQ while our approach gets higher FID. We believe that it is due to compromising fidelity for quality, as observed in examples in the visual results of Wider-Test-200 in the supplementary material. Our fine-tuned model, however, improves on all metrics on this real-world dataset.

4.3. Ablation Study

Effects of fine-tuning dataset size. We investigate the effects of varying fine-tuning dataset size. In Tab. 4, we compare the results of the fine-tuning pre-trained SwinIR fine-tuning datasets of different sizes (i.e. different number of low-quality and pseudo target pairs). We starts to obtain consistent improvements even with only 100 fine-tuning images. The fine-tuned SwinIR becomes worse with 20 fine-tuning images. We believe that the SwinIR has over-fitted to this extremely small fine-tuning dataset in this case, which causes slight performance degradation. The same ablation on fine-tuned CodeFormer is in the supplementary material.

Effects of timesteps and low pass filter choices. In the pseduo target generation stage, our approach involves applying a forward diffusion process on low-quality restoration to timestep $t = K$, and then utilizing low frequency constrained denoising process from timestep $t = K$ down to $t = L$, followed by the standard unconditional denoising for the rest of the timesteps. Having downsampling factor of $N = 8$ for the low pass filter gives better fidelity to the inputs, but with artifacts carried over, while larger N produces targets with higher perceptual quality. Overall, $N = 16$ gives the best balance between fidelity and quality (Fig. 6 (a)). As shown in Fig. 6 (b), applying low frequency constraint down to $L = 180$ and to $L = 0$ would also produce lower quality images due to artifacts in the low frequency content at small timesteps, while our approach is robust to the starting timestep ($K = 900$). We also show the results of unconditional denoising for all timesteps in (c), where small timesteps of ($K = 180$) and ($K = 360$) produces artifacts, and a large timestep of ($K = 600$) distorts the image content due to the lack of regularization during denoising.

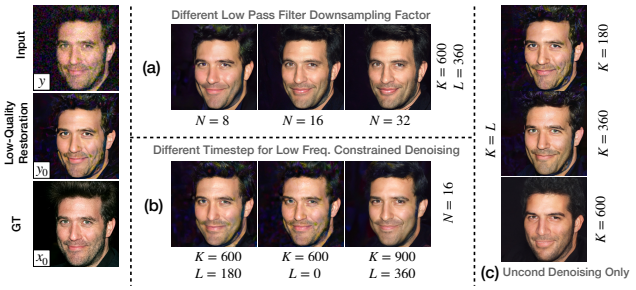

Figure 6. **Effects of low pass downsampling factor and timestep choices on pseudo target.** In (a) we show the effects of adjusting the low pass filter downsampling factor (N); In (b) we show the effects of different timestep windows for low frequency constrained denoising (K and L); In (c) we show the pseudo targets if only unconditional denoising is applied (**zoom in for details**). A large version of this figure is included in the supplementary material for better visualization of the details.

5. Conclusion

In this paper, we propose an unsupervised approach for blind face restoration that addresses the problem of pre-trained restoration models failing on out-of-distribution degradations. Our method requires neither paired ground-truth high-quality images nor knowledge of the inputs' degradation process. Our two-stage pipeline starts by generating pseudo targets using a pre-trained diffusion model with a combination of low frequency constrained denoising and unconditional denoising. We then fine-tune the pre-trained models with pairs of low-quality inputs and pseudo targets. Our approach can consistently improve a pre-trained model's performance on out-of-distribution degradations, producing realistic restoration with satisfactory balance of fidelity and perceptual quality. Our best fine-tuned model achieves state-of-the-art blind face restoration on both synthetic and real-world datasets.

References

[1] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 2

[2] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*, 2021. 2

[3] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 2

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021. 2, 4, 5

[5] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *CVPR*, 2023. 2

[6] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023. 2

[7] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, 2022. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 5

[9] Zheng Ding, Xuaner Zhang, Zhuowen Tu, and Zhihao Xia. Restoration by generation with constrained priors. In *CVPR*, 2024. 1, 4

[10] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 2

[11] Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *ICLR*, 2024. 2

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[13] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, 2023. 2

[14] Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *ICCV*, 2023. 2

[15] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. In *TIP*, 2008. 5

[16] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In *CVPR*, 2023. 2, 4

[17] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *CVPR*, 2023. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014. 5

[19] Yuchao Gu et al. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 2022. 1, 2, 7

[20] 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 *CVPR*, 2023. 2

[21] 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 *NeurIPS*, 2017. 6

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv*, 2017. 5

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 5, 6

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[26] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022. 2

[27] Bahjat Kawar, Jiaming Song, Stefano Ermon, and Michael Elad. Jpeg artifact correction using denoising diffusion restoration models. In *NeurIPS Workshop*, 2022. 2

[28] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. In *NeurIPS*, 2021. 2

[29] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 6

[30] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv*, 2019. 2

[31] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020. 1, 2, 5

[32] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, 2020. 2

[33] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 1, 2, 5

[34] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. In *TPAMI*, 2022. 2

[35] Zelin Li, Dan Zeng, Xiao Yan, Qiaomu Shen, and Bo Tang. Analyzing and combating attribute bias for face restoration. In *AAAI*, 2023. 2

[36] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 4, 5, 6, 7

[37] 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*, 2023. 1, 2, 7

[38] Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Practical signal-dependent noise parameter estimation from a single noisy image. In *TIP*, 2014. 5

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2

[40] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. In *ICML*, 2023. 2

[41] 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 *CVPRW*, 2023. 2

[42] Markku Makitalo and Alessandro Foi. Optimal inversion of the generalized anscombe transformation for poisson-gaussian noise. In *TIP*, 2012. 5

[43] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 5

[44] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 2

[45] Naoki Murata, Koichi Saito, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Gibbsddrm: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *ICML*, 2023. 2

[46] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. In *TPAMI*, 2023. 2

[47] Xinmin Qiu, Congying Han, ZiCheng Zhang, Bonan Li, Tiande Guo, and Xuecheng Nie. Diffbfr: Bootstrapping diffusion model towards blind face restoration. *arXiv*, 2023. 2

[48] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *ICCV*, 2023. 2

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[51] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *NeurIPS*, 2023. 2

[52] Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv*, 2023. 2

[53] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 2

[54] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. In *TPAMI*, 2022. 2

[55] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 3

[56] Donghwan Seo, Abhijith Punnappurath, Luxi Zhao, Abdelrahman Abdelhamed, Sai Kiran Tedla, Sanguk Park, Jihwan Choe, and Michael S Brown. Graphics2raw: Mapping computer graphics images to sensor raw images. In *ICCV*, 2023. 5

[57] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 2

[58] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 3

[60] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023. 2

[61] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[62] Xintao Wang et al. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 1, 2, 5, 7

[63] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023. 2

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[65] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *CVPR*, 2022. 1, 2

[66] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *CVPR*, 2023. 1, 2, 4, 5, 7

[67] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *CVPR*, 2022. 2

[68] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *ICCV*, 2023. 2

[69] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. In *NeurIPS*, 2023. 1, 2, 7

[70] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. 6

[71] 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*, 2022. 6

[72] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 1, 2

[73] Rajeev Yasarla, Federico Perazzi, and Vishal M Patel. Deblurring face images using uncertainty guided multi-stream semantic networks. In *TIP*, 2020. 2

[74] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, 2018. 2

[75] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *arXiv*, 2022. 1, 2, 4, 5, 7

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6

[77] Yang Zhao, Tingbo Hou, Yu-Chuan Su, Xuhui Jia, Yandong Li, and Matthias Grundmann. Towards authentic face restoration with iterative diffusion models and beyond. In *ICCV*, 2023. 2

[78] Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu, Changyou Chen, and Xuhui Jia. Rethinking deep face restoration. In *CVPR*, 2022. 2

[79] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 1, 2, 5, 6, 7

[80] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *CVPR*, 2022. 2

[81] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, 2016. 2

[82] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *CVPRW*, 2023. 2