# Unlocking the Potential of Diffusion Priors in Blind Face Restoration

Yunqi Miao[1] Zhiyu Qu[3] Mingqi Gao[2] Changrui Chen[1]
Jifei Song[3] Jungong Han[2] Jiankang Deng[4]
[1]University of Warwick [2]University of Sheffield [3]University of Surrey [4]Imperial College London
{yunqimiao709, im.mingqi, jungonghan77, jiankangdeng}@gmail.com, changrui.chen@warwick.ac.uk, z.qu@surrey.ac.uk

## Abstract

*Although diffusion prior is rising as a powerful solution for blind face restoration (BFR), the inherent gap between the vanilla diffusion model and BFR settings hinders its seamless adaptation. The gap mainly stems from the discrepancy between 1) high-quality (HQ) and low-quality (LQ) images and 2) synthesized and real-world images. The vanilla diffusion model is trained on images with no or less degradations, whereas BFR handles moderately to severely degraded images. Additionally, LQ images used for training are synthesized by a naive degradation model with limited degradation patterns, which fails to simulate complex and unknown degradations in real-world scenarios. In this work, we use a unified network FLIPNET that switches between two modes to resolve specific gaps. In restoration mode, the model gradually integrates BFR-oriented features and face embeddings from LQ images to achieve authentic and faithful face restoration. In degradation mode, the model synthesizes real-world like degraded images based on the knowledge learned from real-world degradation datasets. Extensive evaluations on benchmark datasets show that our model 1) outperforms previous diffusion prior based BFR methods in terms of authenticity and fidelity, and 2) outperforms the naive degradation model in modeling the real-world degradations.*

## 1. Introduction

Blind face restoration (BFR) aims to recover high-quality (HQ) facial images from various degradations. The ill-posed nature of this task stems from the complexity and uncertainty of degradations in real-world scenarios. Versatile facial priors have been used to help BFR, such as facial landmarks [4], referential HQ images or features [31, 35], and the pretrained StyleGAN [33, 40]. The success of diffusion models (DMs) [7, 28] in generation tasks has recently excited the restoration field, as their ability to generate HQ content from noise aligns perfectly with the restoration objective [21, 22, 36, 38]. Despite its powerful generative ability, the inherent gap between the vanilla diffusion process and the BFR setting hinders its seamless adaptation. The gap arises mainly from the distribution mismatch between: 1) HQ images and low-quality (LQ) images and 2) synthesized images and real-world images.

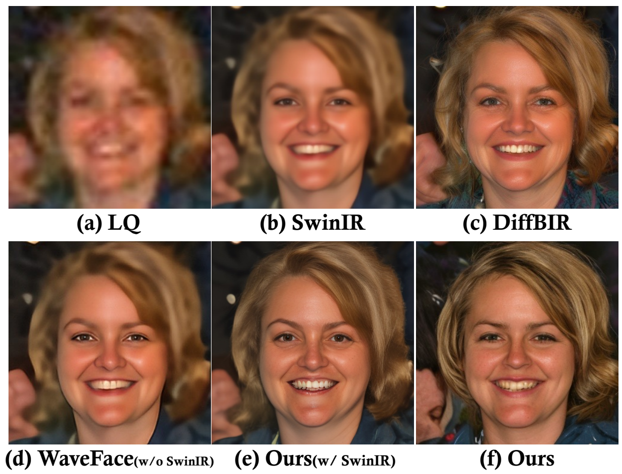


Figure 1. Restoration results of WIDER [39]. Pre-processing module (*i.e.*, SwinIR) based methods (*i.e.*, DiffBIR) fail to render fine-grained details. Methods with condition-concatenated inputs (*i.e.*, WaveFace) unable to handle unknown and complex degradation in real-world scenarios, leading to over-smoothed results.

The vanilla DMs are trained on natural images with minor or no degradation, making them unfit the BFR task, which primarily targets moderately to severely degraded face images. To bridge the gap, previous attempts [21, 23] brutally remove the severe degradations via a pre-processing module [20], followed by a DM to render facial details. However, the pre-processing module also removes discriminative details such as wrinkle and freckle (Fig. 1(b)), thus compromising the fidelity of restoration (Fig. 1(c)). Some approaches [22] bypass the pre-processing module by concatenating the degraded images, acting as the condition, to HQ inputs. Although the first gap is mitigated by gradually introducing LQ features throughout denoising, they fail to deliver adequate results on real-world data (Fig. 1(d)) as the degradation model used for training set synthesizing fails to simulate real-world degradations. A pioneering work [34] applies the naive degradation model repeatedly to HQ images to create LQ counterparts. Although effective, this work generates less diverse data due to its reliance on limited degradation patterns, which is incompatible with "blind" face restoration.

To address these problems, this work uses only a large-scale Text-to-Image (T2I) model with few trainable parameters, eliminating the reliance on pre-processing modules in both training and inference. Our goal is to steer the pretrained T2I model towards authentic and faithful face restoration while preserving its strong image prior by properly integrating contributive features from LQ images. Despite offering valuable cues such as color and structure for faithful face restoration, LQ images also carry the degradation that affects authenticity. To strike a balance, we propose BoostHub to selectively integrate BFR-oriented features from LQ inputs while discarding irrelevant ones. Meanwhile, we identify the misalignment between face embeddings used in personalized T2I generation [24, 25] and those needed for BFR, which inspires us to explore BFR-oriented face embeddings. Apart from face restoration, we unlock the potential of diffusion prior in modeling real-world degradations. Being trained with HQ and degraded image pairs collected from real-world scenarios, the T2I model learns the distribution of real-world degradations, which is used to synthesize degraded images.

In this work, we unify the above solutions into a single model, *i.e.*, FLIPNET, which takes HQ and LQ image pairs as input. By simply flipping the inputs, the model switches between restoration mode and degradation mode, achieving authentic and faithful face restoration and real-world degradation synthesis, respectively. The contributions of this work can be summarized as follows:

- A two-pronged solution, *i.e.*, FLIPNET, is proposed to bridge two inherent gaps that hinder the seamless adaptation of diffusion prior in BFR by simply switching between restoration mode and degradation mode.
- When switching to restoration mode, FLIPNET achieves authentic and faithful face restoration by integrating BFR-oriented features and face embeddings from LQ images.
- When switching to degradation mode, FLIPNET learns from degraded images in real-world scenarios and, in turn, synthesizes such images with acquired knowledge.
- Comprehensive experiments demonstrate the superiority of FLIPNET in integrating diffusion prior to BFR.

## 2. Related works

**Blind face restoration (BFR).** BFR works are generally based on three priors: 1) **Geometric priors**, including facial landmarks [4, 6] and facial parsing maps [3, 43], are used to provide explicit facial structure information. They often yield inferior restoration results due to the imprecise prediction on degraded images. 2) **Reference priors** based methods either leverage high-quality (HQ) reference images to provide identity information [32] or construct a dictionary that restores HQ facial features [31, 35]. But the restoration performance is restricted by the authenticity of reference images and the size of the codebook. 3) **Generative priors** encapsulated in StyleGAN2 [14] are used to provide rich facial details for restoration. GFP-GAN [33] and GPEN [40] restore degraded faces by learning the mapping from LQ input to the latent space where GAN prior network can reproduce the desired HQ face images.

**Diffusion prior in BFR.** Inspired by the ability of diffusion models (DMs) [28, 30] in emerging HQ content from noise, diffusion prior has become a favorable solution for BFR. However, DMs are typically trained on images with less or no degradation while BFR deals with moderately to severely degraded images, which makes its seamless adaptation challenging. Therefore, some works break BFR down to two stages: degradation removal and details generation. DR2 [36] employs the diffusion prior at first stage to obtain degradation-invariant images, where HQ counterparts are obtained by the follow-up face restoration model. DiffBIR [21] and PMRF [23] first coarsely remove severe degradations with a pretrained pre-processing module [20]. Given the obtained intermediate results, DiffBIR renders realistic facial details via a follow-up T2I model. Whilst PMRF takes such images as the initial state of the distribution transportation, where a rectified flow model is optimized to model the path to HQ counterparts. Instead of adopting such a two-stage scheme, PGDiff [38] uses a pretrained BFR model [34] to provide instructive semantics. Throughout the denoising process, the model is trained to minimize the discrepancy between intermediate outputs and target properties. WaveFace [22] uses diffusion prior to restore the low-frequency component exclusively, where low-frequency subbands of LQ images are concatenated to noise and serve as condition for generation. Although exciting restoration results have been achieved, the pre-processing module will also smooth out facial details, thereby compromising the fidelity. Additionally, the reliance on other face restoration models imposes limitations on the potential of the proposed model. How to effectively unleash the potential T2I model in BFR task has NOT been fully exploited.

**Real-world degradation modeling.** Apart from the gap between HQ and LQ images, diffusion prior BFR methods general deliver inferior performance on real-world images as the degradation model used to synthesize training set fail to simulate that in real-world scenarios. Classical degradation model [18] includes four types of degradations: blur, noise, resize, and JPEG compression. The simple combination fails to handle real-world cases where images typically spread several times over different digital devices. Thus, Real-ESRGAN [34] proposes a "high-order" degradation model, where classical degradations with different hyper-parameters are applied repeatedly on HQ images to synthesize the degraded counterparts. While this may serve as a stopgap, the synthesis remains constrained by the limited degradation patterns. How to effectively simulate complex and blind degradations remains a formidable challenge.

In this work, we address above challenges with a unified model that is solely based on a large-scale T2I model with few trainable parameters. The model provides a two-pronged solution to both authentic and faithful face restoration and real-world degradation modeling.

## 3. Preliminary

**Stable Diffusion** is a large-scale text-to-image (T2I) latent diffusion model, which performs both diffusion and denoising processes in latent space. Specifically, an autoencoder [16] converts an image $x$ to a latent $z$ with encoder $\mathcal{E}$ and reconstructs it with decoder $\mathcal{D}$, *i.e.*, $\hat{x} = \mathcal{D}(\mathcal{E}(x))$. In diffusion process, Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ is applied to the starting latent $z_0$ to achieve noisy latent $z_t$:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \tag{1}$$

where $\{\bar{\alpha}_t\}_{t=1}^T$ is a pre-defined noise variance schedule over timestep $t$. For denoising, a network $\epsilon_\theta$ is optimized to predict the noise $\epsilon$ at a particular timestep $t$ conditioned on $c$, such as text prompt, which follows the objective:

$$\mathcal{L}_{ldm} = \mathbb{E}_{z, \epsilon \sim \mathcal{N}(0,1), t} \Big[ \| \epsilon - \epsilon_\theta(z_t, t, c) \|_2^2 \Big]. \tag{2}$$

Given the noise predicted at timestep $t$, the denoised latent $\hat{z}_0$ can be predicted by:

$$\hat{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t, c) \right). \tag{3}$$

The denoised image can be further obtained by $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$.

**Low-Rank Adaptation (LoRA)** is widely used for the efficient adaptation of large language models or text-to-image models to downstream tasks [10]. Concretely, two low-rank matrices $B \in \mathbb{R}^{n \times r}$ and $A \in \mathbb{R}^{r \times m}$ are introduced to update the base model weights $W_0 \in \mathbb{R}^{n \times m}$ with $\triangle W = BA$, where $r \ll min(n, m)$ refers to the intrinsic rank of $\triangle W$. The updated weight matrix $W = W_0 + \triangle W$ is used for inference. By updating $A$ and $B$ exclusively during training, the base model gradually adapts to the specific domains. Meanwhile, since $W_0$ is frozen throughout the optimization, the generative ability of base model is well preserved. For efficiency, LoRA is typically applied only to attention layers for model fine-tuning.

## 4. Methodology

This work aims to handle inherent gaps between the vanilla diffusion prior and blind face restoration (BFR) setting, which stem from the distribution mismatch between 1) high-quality (HQ) and low-quality (LQ) images, and 2) synthesized and real-world images. Unlike previous attempts addressing the first gap with pre-processing modules [21, 23] or guidance from other face restoration methods [38], our model, namely FLIPNET, is simply built upon a T2I model with few trainable parameters. The framework is illustrated in Fig. 2, which takes pairs of HQ and LQ images as input. By simply flipping the input, the model switches between restoration mode (Sec. 4.1) and degradation mode (Sec. 4.2) to achieve authentic and faithful face restoration and synthesizing real-world degraded images, respectively. Note that, to preserve the strong image prior of the large-scale T2I model, we fine-tune only the LoRA weights [10] plugged to all attention layers.

### 4.1. Restoration mode

In restoration mode, the denoising network takes HQ images $\boldsymbol{x}$ as input and LQ images $\boldsymbol{y}$ as condition. Unlike previous works concatenating LQ images [22] or pre-processed results [21, 23] to HQ images, our model is conditioned on diffused LQ images. Specifically, the diffusion process (Eq. (1)) is applied to both HQ and LQ images, where noise ($\epsilon_x$, $\epsilon_y$) and timestep ($t_x$, $t_y$) are sampled separately. The noisy pairs ($\boldsymbol{x}_{t_x}, \boldsymbol{y}_{t_y}$) are concatenated along the batch axis and fed to the denoising network, which is optimized to simultaneously denoise both inputs, enabling the restoration of HQ images from LQ counterparts under varying noise levels. However, LQ images act as a double-edged sword, which improves the restoration fidelity by providing structural information while compromising the authenticity. To strike a balance, BoostHub is proposed to selectively integrate BFR-oriented features, and meanwhile, discard irrelevant ones. Additionally, we identify the misalignment between face embeddings widely-used for personalized generation [24, 25] and those required by BFR, which inspires us to explore BFR-oriented face embeddings.

**BoostHub.** Following the pioneering T2I model [28], our denoising network $\epsilon_\theta$ is implemented as UNet, with its blocks comprising ResNet layers, self-attention layers, and cross-attention layers. Inspired by a joint module used in conditioned generation [19], we use BoostHub to enhance noisy HQ features with BFR-oriented features from LQ conditions. As shown in Fig. 2, BoostHub is placed in parallel to self-attention layers to involve LQ features without interrupting the spatial relationships within HQ images.

As illustrated in Fig. 2, given the noisy HQ and LQ features ($F_x$, $F_y$), BoostHub projects them to query, key and value via projection matrices $W_{Q_x}$, $W_{K_y}$ and $W_{V_y}$ and then integrate them with attention mechanism:

$$\begin{aligned} F_{ro} &= W_O \cdot \text{Attn}(Q_x, K_y, V_y) \\ &= W_O \cdot \text{Softmax}(Q_x K_y^T / \sqrt{d}) V_y, \end{aligned} \tag{4}$$

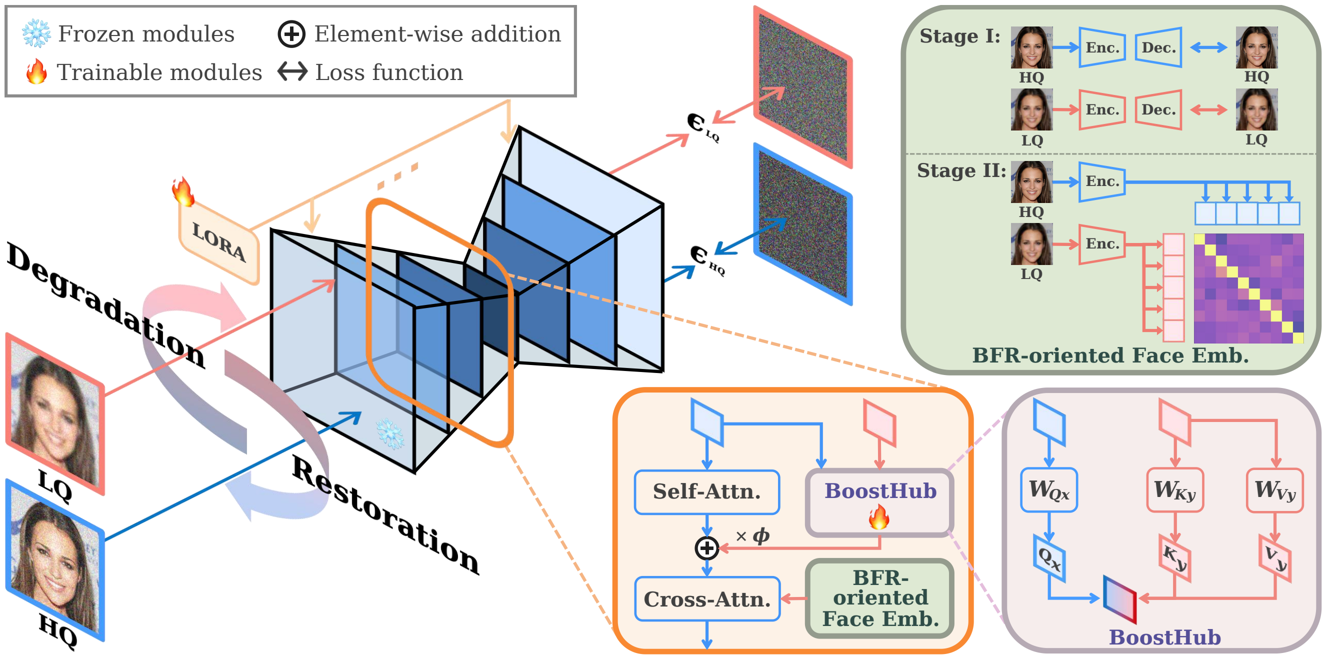


Figure 2. **FLIPNET pipeline.** FLIPNET takes noisy high-quality (HQ) and low-quality (LQ) image pairs as input. It can switch between restoration mode and degradation mode. Taking restoration mode as an example, HQ images are used as input while LQ images are condition, where a BoostHub is placed in parallel to each self-attention layer to selectively integrate the BFR-oriented LQ features for denoising. Additionally, LoRA weights are plugged to all self-attention and cross-attention layers to adapt the base model to face domain. By simply flipping the image-condition order, it switches to degradation mode to synthesize degraded data with real-world degradations.

where $Q_x = W_{Q_x} F_x$, $K_y = W_{K_y} F_y$, and $V_y = W_{V_y} F_y$. Following prior arts [8], $W_{Q_x}$, $W_{K_y}$ and $W_{V_y}$ are initialized by corresponding self-attention weight of the base model and the output projection matrices $W_O$ are initialized by all-zero matrix to avoid harmful effects that additional modules might introduce. The BFR-oriented features $F_{ro}$ are then combined with self-attention output $F_{sa_x}$ as:

$$F_{joint} = F_{sa_x} + \phi \cdot F_{ro} = \text{Self-Attn}(F_x) + \phi \cdot F_{ro}, \quad (5)$$

where the enhance weight $\phi$ is set to control how much LQ features will be involved. Subsequently, the joint features $F_{joint}$ are integrated with text prompts as well as BFR-oriented face embeddings via cross-attention.

**BFR-oriented face embeddings.** In personalized generation works [24, 25], face embeddings, acting as an additional condition, are integrated via cross-attention layers to control the identity. However, prior arts generally leverage ID embeddings from a face recognition network [5]. Yet, such embeddings are trained to distinguish different persons instead of preserving fine-grained facial details, making them ineffective in providing instructive guidance for restoration. To this end, this work aims to explore BFR-oriented face embeddings, which should 1) well represent the appearance of the given person, and 2) be positioned in the common latent space of both HQ and LQ inputs so that they will not be biased when integrating with joint features from BoostHub in cross-attention layers. Inspired by the training scheme used in DAEFR [31], we propose to obtain such embeddings from an autoencoder [16] trained for face reconstruction. Instead of the discrete codebook, we adopt continuous latent features which can provide fine-grained details. A two-stage learning scheme is adopted.

**a) Reconstruction stage.** First, HQ/LQ autoencoders are trained to reconstruct input face images. HQ/LQ images $\boldsymbol{x}/\boldsymbol{y} \in \mathbb{R}^{H\times W\times 3}$ are projected to latent space $\boldsymbol{z}_{x/y} \in \mathbb{R}^{h\times w\times d}$ via encoder and then projected back to images $\hat{\boldsymbol{x}}/\hat{\boldsymbol{y}}$ by decoder. To achieve faithful reconstruction, the objective consists of L1 loss $\mathcal{L}_1$, perceptual loss $\mathcal{L}_p$, and adversarial loss $\mathcal{L}_{adv}$, *i.e.*, $\mathcal{L}_{ae} = \mathcal{L}_1 + \lambda_{ap} \cdot \mathcal{L}_p + \lambda_{adv} \cdot \mathcal{L}_{adv}$, where $\lambda_{ap}$ and $\lambda_{adv}$ are set as 0.5 and 0.8, respectively.

**b) Association stage.** Similar to CLIP [26], we fine-tune the autoencoders via the cross-entropy loss to map HQ/LQ features into a common latent space. HQ and LQ latents are first flattened to $\{\boldsymbol{z}^i_{x/y} \in \mathbb{R}^d\}^N_{i=1}, N = h \times w$. The similarity matrix between $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$ can be computed as $\mathcal{M} \in \mathbb{R}^{N\times N}$. We adopt the cross-entropy loss ($\mathcal{L}^H_{ce}$ and $\mathcal{L}^L_{ce}$) to maximize the similarity between corresponding patches via maximizing scores in the diagonal:

$$\mathcal{L}^{H(L)}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{i,j}\log(z^{i,j}_{x(y)}), \quad (6)$$

where $y_{i,j}$ represents ground-truth labels. The objective of the association stage is $\mathcal{L}_{asso} = \mathcal{L}_{ae} + (\mathcal{L}^H_{ce} + \mathcal{L}^L_{ce})/2$.

Once trained, the LQ encoder is used to extract face embeddings $\boldsymbol{z}_y \in \mathbb{R}^{N\times d}$, which are then passed through a lightweight adapter $\tau(\cdot)$ to align with the dimensionality of text tokens before integrating with the T2I model, *i.e.*, $c_{\text{id}} = \tau(\boldsymbol{z}_y)$. The adapter consists of several linear layers with Layer Normalization [41].

**Objectives.** In restoration mode, the model learns how to denoise the noisy $(\boldsymbol{x}_{t_x}, \boldsymbol{y}_{t_y})$ jointly following the objective:

$$\mathcal{L}_{ldm} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\epsilon,t_x,t_y}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{x}_{t_x}, \boldsymbol{y}_{t_y}, t_x, t_y, c)\|^2\right], \quad (7)$$

where $\epsilon = (\epsilon_x, \epsilon_y)$ and $c = (c_{\text{text}}, c_{\text{id}})$. Beyond latent-level constraints, image-level constraints are applied to the denoised predictions $\hat{\boldsymbol{x}}_0$ at each timestep obtained by Eq. (3), which include MSE loss $\mathcal{L}_{mse}$ and perceptual loss $\mathcal{L}_p$:

$$\mathcal{L}_{mse} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}_0\|_2^2, \quad \mathcal{L}_p = \|\phi(\boldsymbol{x}) - \phi(\hat{\boldsymbol{x}}_0)\|_2^2, \quad (8)$$

where $\phi(\cdot)$ represents VGG19 [29] backbone. The overall training objective is $\mathcal{L}_{rm} = \mathcal{L}_{ldm} + \lambda_{mse} \cdot \mathcal{L}_{mse} + \lambda_p \cdot \mathcal{L}_p$. $\lambda_{mse}$ and $\lambda_p$ are set as 1 and 0.01 throughout the training.

### 4.2. Degradation mode

By simply flipping the input, *i.e.*, LQ images $\boldsymbol{y}$ serving as input while HQ images $\boldsymbol{x}$ as condition, our FLIPNET switches to degradation mode. To learn the real-world degradations, our model needs to be trained with paired HQ and LQ images collected from real-world scenarios. In the absence of such face datasets, we gather multiple low-level vision datasets captured in real-world scenarios to construct the training dataset for degradation mode, including Dense-Haze [2] for image dehazing, LOL [37] for low light enhancement, SIDD [1] for image denoising, and RealBlur [27] for image deblurring. To construct the training set, full set of Dense-Haze and LOL are included and 1000 pairs are randomly selected from SIDD and RealBlur-J, obtaining a real-world dataset with around 2500 pairs. To keep the model's ability in synthesizing the classical degradations Eq. (9), 4000 face images are randomly selected from FFHQ to synthesize LQ counterparts on-the-fly during training. The training process is the same with restoration mode, except for the condition and the objective.

**Cross-attention conditions.** Being trained on both face and non-face datasets, we do NOT leverage BFR-oriented face embeddings in this mode. In other words, text prompts are the only condition for cross-attention layers, which are generated by BLIP [17] following previous works [8].

**Objectives.** Only diffusion loss $\mathcal{L}_{ldm}$ is used, *i.e.*, $\mathcal{L}_{dm} = \mathcal{L}_{ldm}$, as imposing image-level constraints on degraded images will introduce unforeseen noise to predictions.

Once trained, the model can be used to synthesize degraded counterparts of HQ face images with the acquired knowledge. The generated images carry diverse and complex degradations, with the intensity of degradation remaining entirely unknown, which are more suitable for "BLIND" face restoration. The visualization and distribution of degraded images synthesized by a prior art [34] and our method are illustrated in Fig. 5. As observed, FLIPNET could well model real-world degradations, evidenced by its diverse distribution across multiple real-world face datasets. The synthesized degraded images can be subsequently used to train our FLIPNET under restoration mode.

## 5. Experiments

### 5.1. Datasets and Settings

**Training Dataset.** FFHQ [13], containing 70k high-quality (HQ) face images, is used as the training set. Following previous works [33, 36], images are resized to $512 \times 512$ before synthesis. Two degradation models are adopted to synthesize low-quality (LQ) counterparts:

**1) On-the-fly degradation.** Considering its efficiency, Real-ESRGAN [34] is used to synthesize degraded images online following the formula:

$$\boldsymbol{y} = \left\{[(\boldsymbol{x} \otimes \boldsymbol{k}_\sigma)\downarrow_s + \boldsymbol{n}_\delta]_{\text{JPEG}_q}\right\}\uparrow_s, \quad (9)$$

where a HQ image $\boldsymbol{x}$ is firstly blurred by a Gaussian kernel $\boldsymbol{k}_\sigma$, followed by a downsampling of scale $s$. Afterward, Gaussian noise $\boldsymbol{n}_\delta$ and JPEG compression with quality factor $q$ are applied to the image, which is then upsampled back to the original size to obtain its LQ counterpart $\boldsymbol{y}$. The hyper-parameters $\sigma$, $s$, $\delta$, and $q$ are uniformly sampled from $[0.1, 15]$, $[0.8, 32]$, $[0, 20]$, and $[30, 95]$ respectively. On top of this, the process is performed for the second time, where random degradations are applied in shuffled order.

**2) Off-shelf degradation.** In light of its diversity and authenticity, our FLIPNET is switched to degradation mode to synthesize degraded images offline. For each HQ face image, we generate five images with diverse and unknown degradations, serving as LQ counterparts.

For each iteration, degraded images generated by on-the-fly and off-shelf schemes are selected by a probability of 0.5 to construct a training batch.

**Testing Dataset.** We evaluate FLIPNET on a synthetic dataset: CelebA-Test and three real-world datasets: LFW [11], WebPhoto [33] and WIDER [44]. CelebA-Test contains 3000 HQ images from CelebA-HQ [12]. Akin to the training set, LQ counterparts in test set are synthesized by both on-the-fly and off-shelf degradation scheme at equal probability. In terms of three real-world datasets, LFW contains 1711 mildly degraded face images in the wild, which comprises the first image for each person in LFW [11]. WebPhoto includes 407 images crawled from the internet, some of which are old photos with severe degradation. WIDER consists of 970 images with severe degradations from the WIDER dataset [39].

**Evaluation Metrics.** For evaluation, we adopt two pixel-wise metrics (PSNR and SSIM), a reference perceptual met-

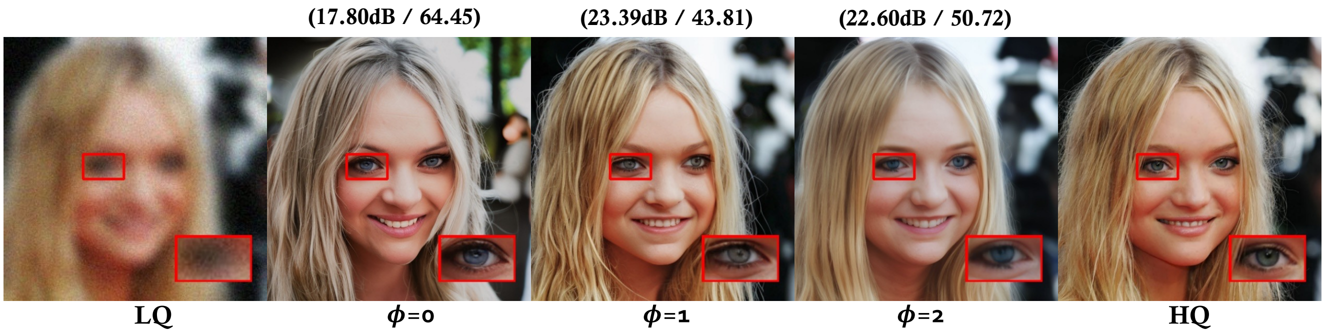


Figure 3. Visualization of the effectiveness of enhance weight $\phi$ on CelebA-Test. $\phi$ strikes a balance between fidelity and authenticity by determining the amount of LQ features involved. PSNR(↑) / Deg.(↓) are reported.

ric (LPIPS [42]), and a non-reference perceptual metric (FID [9]). To measure the consistency of identity, the angle between ID embeddings extracted by ArcFace ("Deg.") [5] is used. All metrics are used for the evaluation of synthetic dataset while only FID is used for real-world datasets due to the lack of ground-truth HQ images.

**Implementation Details.** We adopt Stable Diffusion 2.1-base[1] as the base T2I model and plug LoRA modules to all self-attention and cross-attention layers. We train LoRA for 90k iterations (batch size=16), with the intrinsic rank set to 64. The text prompts are generated by BLIP [17] for all datasets. We use AdamW [15] optimizer with the learning rate 5e-5. The training process is conducted on $512 \times 512$ resolution with 8 NVIDIA V100 GPUs.

For restoration mode, autoencoders are initialized by the pretrained weight [28]. We trained model for 200k at reconstruction stage and 80k at association stage (batch size=32). For non-face datasets, the model takes no embeddings other than text prompts as condition. During training, we use a probability of 0.5 to drop text prompt and 0.3 to drop face embeddings to enhance the robustness. For inference, we use words such as "low quality", "blurry", and "poorly rendered hands" as the negative prompt for restoration mode while use null text prompt for degradation mode. The classifier-free guidance (CFG) scale is set as 1.5.

## 5.2. Ablation Studies

### 5.2.1. Restoration mode

**BoostHub enhance weight.** According to Eq. (5), the enhance weight $\phi$ determines the amount of BFR-oriented features integrated from LQ inputs. When less LQ features are involved, the ability of the T2I model in generating high-quality contents is fully exploited while features that facilitate face restoration, such as color and structure, are ignored, thereby affecting the authenticity. Conversely, an over-injection of such features will compromise the restoration quality. Restoration results with different enhance weights as illustrated in Fig. 3.

As observed, without guidance from LQ images ($\phi = 0$), the restoration strictly relies on text prompt and face embeddings. Despite its high quality, the restoration deviates significantly from the original identity. In contrast, LQ features dominate the generation when $\phi = 2$, leading to the poor quality with fine-grained facial details being smoothed out. When we set $\phi = 1$, the model could strike a good balance between fidelity and authenticity.

**BFR-oriented face embeddings.** We identify the misalignment between the widely-used face embeddings from a face recognition model [5] and those required by blind face restoration (BFR). To make a comparison, we visualize the restoration results, conditioned on ArcFace embeddings and our BFR-oriented face embeddings in Fig. 4(a) and Fig. 4(c), as well as corresponding attention maps. As observed, the face recognition model, trained to distinguish different persons, exclusively focuses on features at discriminative regions, such as eyes. As a result, the restoration fails to render fine-grained details at cheeks or hairs.

Beyond presenting facial details, the alignment between HQ and LQ face embeddings is important for our model as face embeddings are supposed to integrated with joint features (Eq. (5)) via cross-attention, which include both HQ features from self-attention layers and BFR-oriented LQ features from BoostHub. The alignment is achieved by "Association stage". To demonstrate its effectiveness, we use face embeddings from the autoencoder trained only with reconstruction stage to perform the restoration. The result is shown in Fig. 4(b). Due to the absence of "Association stage", face embeddings are biased on BFR-oriented LQ features while neglecting HQ ones, leading to compromised the restoration quality. On the contrary, when BFR-oriented face embeddings integrating with the aligned HQ and LQ features, our model is able to deliver faithful results with face details such as wrinkle well preserved (Fig. 4(c)).

### 5.2.2. Degradation mode

**Synthesized images and distribution.** We visualize the degraded images generated by Real-ESRGAN [34] and our

[1] Stable Diffusion v2.1: https://github.com/Stability-AI/stablediffusion

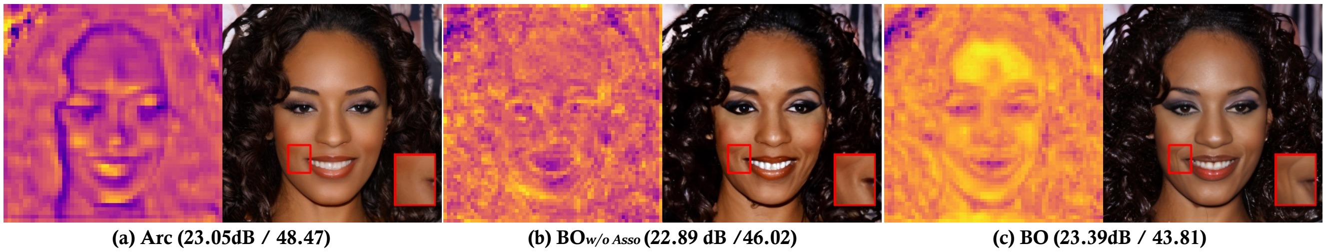


Figure 4. Restoration results (Right) of the model conditioning on face embeddings extracted from ArcFace (Arc), our BFR-oriented ones trained without (BO$_{w/o\ Asso}$) and with (BO) "Association stage" on CelebA-Test. Corresponding attention maps (Left) of embeddings are also shown. PSNR(↑) / Deg.(↓) are reported. The lighter the color, the higher the attention scores.

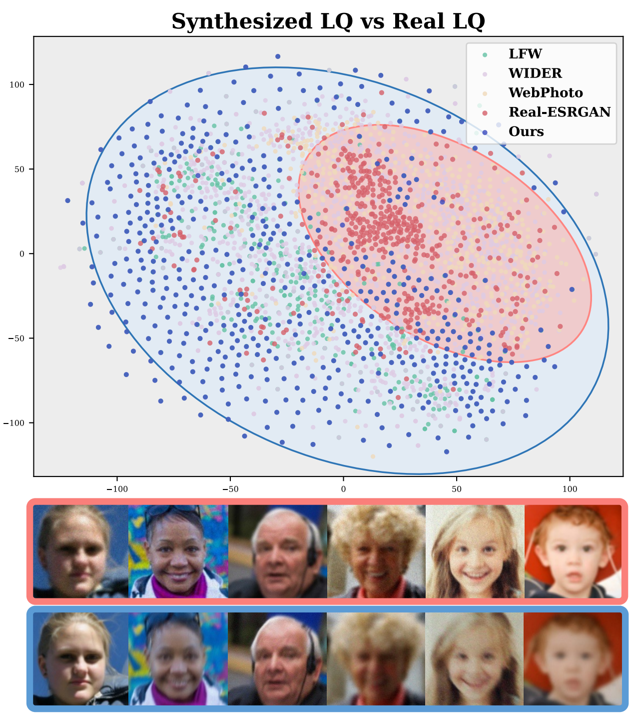


Figure 5. Distribution of degraded images synthesized by Real-ESRGAN [34] and our FLIPNET compared to three real-world datasets: LFW, WebPhoto and WIDER. Synthesized images of **Real-ESRGAN** and **Ours** are shown below.

FLIPNET in Fig. 5. To evaluate how well they fit real-world degradations, the feature distribution of degraded images synthesized by both methods and three real-world datasets: LFW [11], WIDER [44] and WebPhoto [33] is also illustrated. Our model is capable of synthesizing degraded images that are widely distributed across different real-world datasets whereas Real-ESRGAN generates data gathering in a limited region. This can be proved by examples shown below. Images degraded by our model present diverse degradations while that by Real-ESRGAN demonstrate limited degradation patterns.

## 5.3. Comparisons with State-of-the-Art Methods

Evaluations are performed on synthetic and real-world datasets. Unlike previous BFR methods, our training data are synthesized by dual degradation schemes (on-the-fly

Table 1. Quantitative comparison on **CelebA-Test**. "RF++" and "$\mathcal{F}$" refer to "RestoreFormer++" and our FLIPNET. "+O": on-the-fly synthesized images. "+O/F": both on-the-fly and off-shelf synthesized images. "Deg." refers to the angle between identity embeddings of restored images and HQ counterparts. The **best** and the second best results are indicated.

| Methods | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ | Deg.↓ |
|---|---|---|---|---|---|
| GPEN | 23.65/22.86 | 0.61/0.60 | 0.42/0.43 | 18.67/20.20 | 47.24/51.33 |
| GFP-GAN | 23.85/22.70 | 0.61/0.61 | 0.40/0.40 | 17.38/18.52 | 44.96/47.99 |
| RF++ | 24.03/23.15 | 0.63/0.62 | <u>0.38</u>/<u>0.40</u> | 14.98/<u>15.15</u> | 44.00/45.86 |
| DAEFR | 22.70/21.96 | 0.62/0.61 | **0.38**/**0.39** | 14.43/15.65 | 48.44/49.85 |
| DR2 | 22.27/21.83 | 0.63/0.61 | 0.44/0.44 | 31.59/34.49 | 59.24/61.29 |
| PGDiff | 22.87/22.08 | 0.63/0.62 | 0.42/0.42 | 19.48/21.35 | 59.43/60.38 |
| DiffBIR | 24.37/23.39 | 0.62/0.60 | 0.42/0.43 | 20.78/22.54 | 46.58/47.00 |
| WaveFace | 24.16/23.05 | 0.62/0.62 | 0.42/0.43 | 21.44/23.48 | <u>43.80</u>/46.50 |
| PMRF | **24.70**/**23.40** | <u>0.63</u>/<u>0.62</u> | 0.42/0.42 | **13.48**/15.36 | 44.75/<u>45.76</u> |
| $\mathcal{F}_{+O/+O/F}$ | <u>24.44</u>/<u>23.39</u> | **0.64**/**0.63** | 0.42/0.42 | <u>13.82</u>/**15.06** | **43.29**/**43.81** |

Table 2. Quantitative comparisons on **real-world datasets** in terms of **FID**(↓). "RF++" and "$\mathcal{F}$" refer to "RestoreFormer++" and our FLIPNET. "+O": on-the-fly synthesized images. "+O/F": both on-the-fly and off-shelf synthesized images. The **best** and the second best results are indicated.

| Methods | LFW | WebPhoto | WIDER |
|---|---|---|---|
| Input | 124.974 | 170.112 | 199.961 |
| GPEN [40] | 50.792 | 80.572 | 46.340 |
| GFP-GAN [33] | 49.560 | 87.584 | 39.499 |
| RF++ [35] | 50.439 | **75.059** | 49.395 |
| DAEFR [31] | 47.310 | <u>75.453</u> | 35.344 |
| DR2 [36] | 45.298 | 112.344 | 45.348 |
| PGDiff [38] | 44.630 | 89.754 | 36.807 |
| DiffBIR [21] | 44.383 | 91.777 | 35.343 |
| WaveFace [22] | 43.175 | 81.525 | 36.913 |
| PMRF [23] | 50.275 | 81.064 | 40.685 |
| $\mathcal{F}_{+O/+O/F}$ | <u>43.10</u>/**42.98** | 83.34/81.37 | <u>33.09</u>/**31.81** |

and off-shelf). To make a fair comparison, we also synthesize another group with only on-the-fly degradation scheme. The quantitative results are indicated by "+O/F" and "+O" in Tab. 1 and Tab. 2.

**Synthetic dataset.** Quantitative comparisons on CelebA-Test [12] are illustrated in Tab. 1. Methods tend to yield inferior performance on the test set synthesized by dual degradation schemes, as the degradation better aligns with real-world scenarios, in other words, are more challenging.

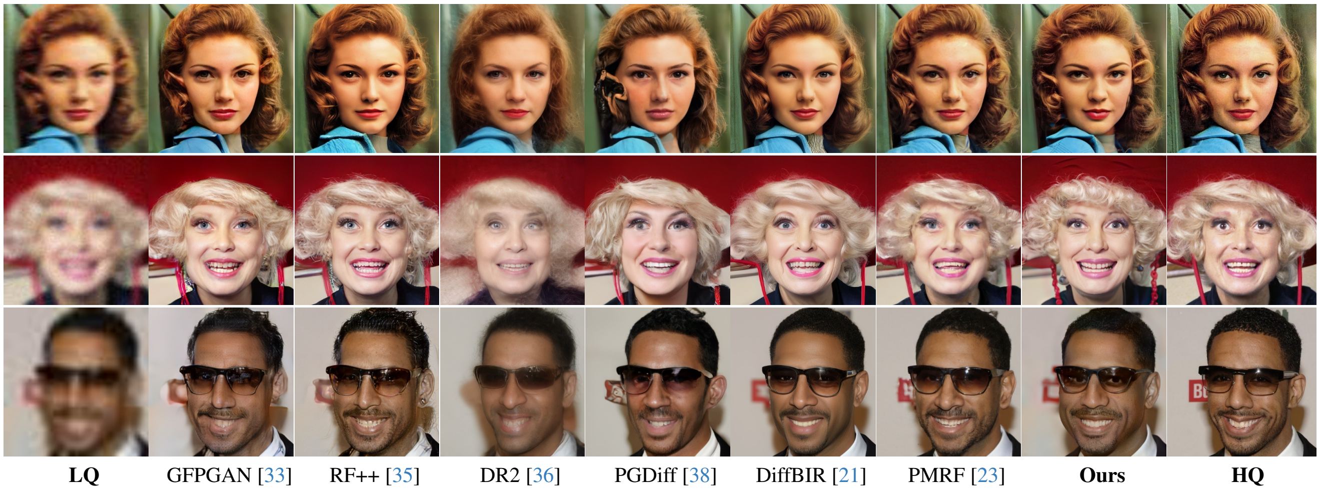


Figure 6. Qualitative comparison with state-of-the-art BFR methods on **CelebA-Test**. Our method achieves authentic and faithful restoration with fine-grained facial details being well preserved. (Zoom in for best view).

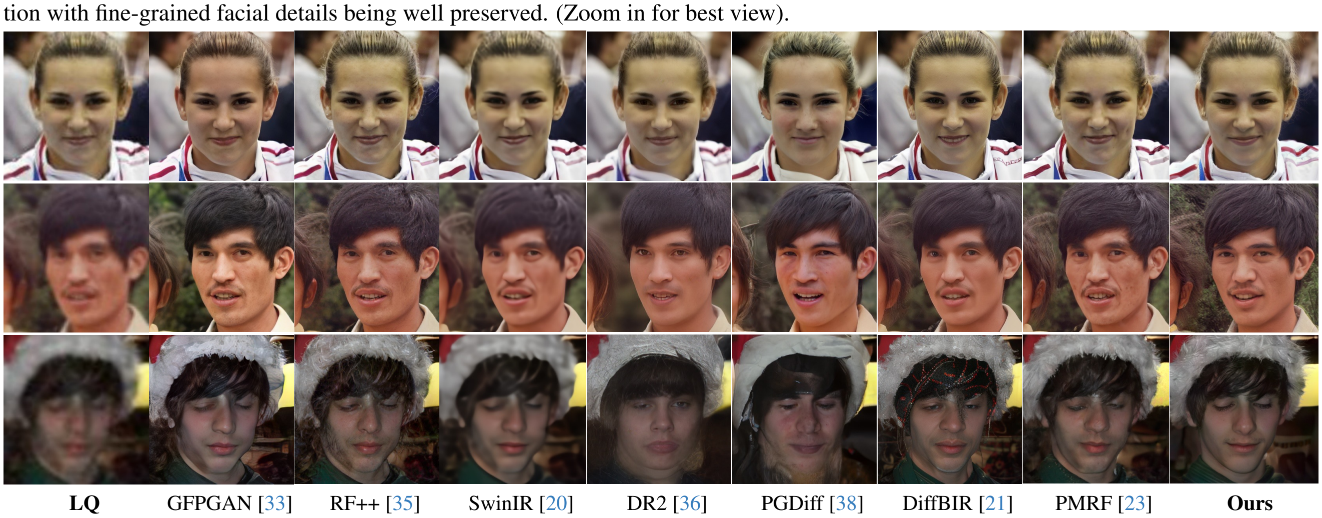


Figure 7. Qualitative comparison with state-of-the-art BFR methods on **real-world** datasets, including LFW (first row), WebPhoto (second row), and WIDER (third row). (Zoom in for best view).

For visualization, we present qualitative comparisons with SOTA methods with increasing degradations in Fig. 6. GFPGAN [33] and RestoreFormer++ [35] struggle with obvious artifacts especially when images are corrupted by mild or severe degradation. They also fail to preserve the identity as the restoration is based on, and also, limited by the StyleGAN prior [33] or the pre-constructed dictionary [35]. As inherent gaps between the vanilla diffusion models and the BFR setting have not been properly handled, previous diffusion prior based methods generally yield unsatisfactory results. DR2 [36] and PGDiff [38] tend to deliver either faithful or authentic results, as the restoration starts from a preliminary result from another face restoration model, making their method suboptimal. DiffBIR [21] and PMRF [23], based on a pre-processing module, typically suffer from rendering vivid facial details or avoiding unexpected artifacts. In comparison, our model delivers authentic and faithful results.

**Real-world datasets.** We show the quantitative comparison on real-world datasets in Tab. 2. FLIPNET outperforms SOTA methods on LFW [11] and WIDER [44]. Qualitative comparisons (Fig. 7) show that previous diffusion prior based methods yield over-smoothed faces and fail to preserve discriminative facial details. For severely-degraded images (Row 3), most methods are plagued by artifacts caused by unknown and complex real-world degradation. In contrast, our approach delivers faithful and authentic restoration results with rich facial details such as mustache.

## 6. Conclusion

We propose FLIPNET to bridge the inherent gap between vanilla diffusion model and BFR settings. The model provides a two-pronged solution, enabling authentic and faithful face restoration and real-world degradations modeling by simply switching between restoration mode and degradation mode.

## References


[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In CVPR, pages 1692–1700, 2018. 5

[2] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In ICIP, pages 1014–1018, 2019. 5

[3] 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

[4] 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. 1, 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019. 4, 6

[6] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. IJCV, 2019. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In NeurIPS, 2021. 1

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In ICLR, 2024. 4, 5

[9] 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

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In ICLR, 2022. 3

[11] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Report, 2008. 5, 7, 8

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In ICLR, 2018. 5, 7

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

[14] 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

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015. 6

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In ICLR, 2014. 3, 4

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, pages 12888–12900, 2022. 5, 6

[18] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In ECCV, pages 272–289, 2018. 2

[19] Xirui Li, Charles Herrmann, Kelvin CK Chan, Yinxiao Li, Deqing Sun, and Ming-Hsuan Yang. A simple approach to unifying diffusion-based conditional generation. In The Thirteenth International Conference on Learning Representations, 2025. 3

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In ICCV, 2021. 1, 2, 8

[21] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. In ECCV, 2024. 1, 2, 3, 7, 8

[22] Yunqi Miao, Jiankang Deng, and Jungong Han. Waveface: Authentic face restoration with efficient frequency recovery. In CVPR, 2024. 1, 2, 3, 7

[23] Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration. In ICLR, 2025. 1, 2, 3, 7, 8

[24] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In ECCV, pages 241–261, 2024. 2, 3, 4

[25] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In CVPR, pages 27080–27090, 2024. 2, 3, 4

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, pages 8748–8763, 2021. 4

[27] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In ECCV, 2020. 5

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022. 1, 2, 3, 6

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015. 5

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv:2010.02502, 2020. 2

[31] Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin CK Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration. In ICLR, 2024. 1, 2, 4, 7

[32] Tuomas Varanka, Tapani Toivonen, Soumya Tripathy, Guoying Zhao, and Erman Acar. Pfstorer: Personalized face restoration and super-resolution. In CVPR, 2024. 2

[33] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In CVPR, 2021. 1, 2, 5, 7, 8

[34] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In ICCV Workshops, 2021. 2, 5, 6, 7

[35] Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. TPAMI, 2023. 1, 2, 7, 8

[36] 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, 5, 7, 8

[37] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In BMVC, 2018. 5

[38] 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, 3, 7, 8

[39] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In CVPR, 2016. 1, 5

[40] 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, 7

[41] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arxiv:2308.06721, 2023. 5

[42] 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. 6

[43] Qingping Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In CVPR, 2022. 2

[44] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In NeurIPS, 2022. 5, 7, 8