DiffBFR: Bootstrapping Diffusion Model for Blind Face Restoration

Xinmin Qiu
University of Chinese Academy of Sciences
Beijing, China
qiuxinmin21@mails.ucas.ac.cn

Congying Han
University of Chinese Academy of Sciences
Beijing, China
hancy@ucas.ac.cn

Zicheng Zhang
University of Chinese Academy of Sciences
Beijing, China
zhangzicheng19@mails.ucas.ac.cn

Bonan Li*
University of Chinese Academy of Sciences
Beijing, China
libonan@ucas.ac.cn

Tiande Guo
University of Chinese Academy of Sciences
Beijing, China
tdguo@ucas.ac.cn

Xuecheng Nie
MT Lab, Meitu Inc.
Beijing, China
nxc@meitu.com

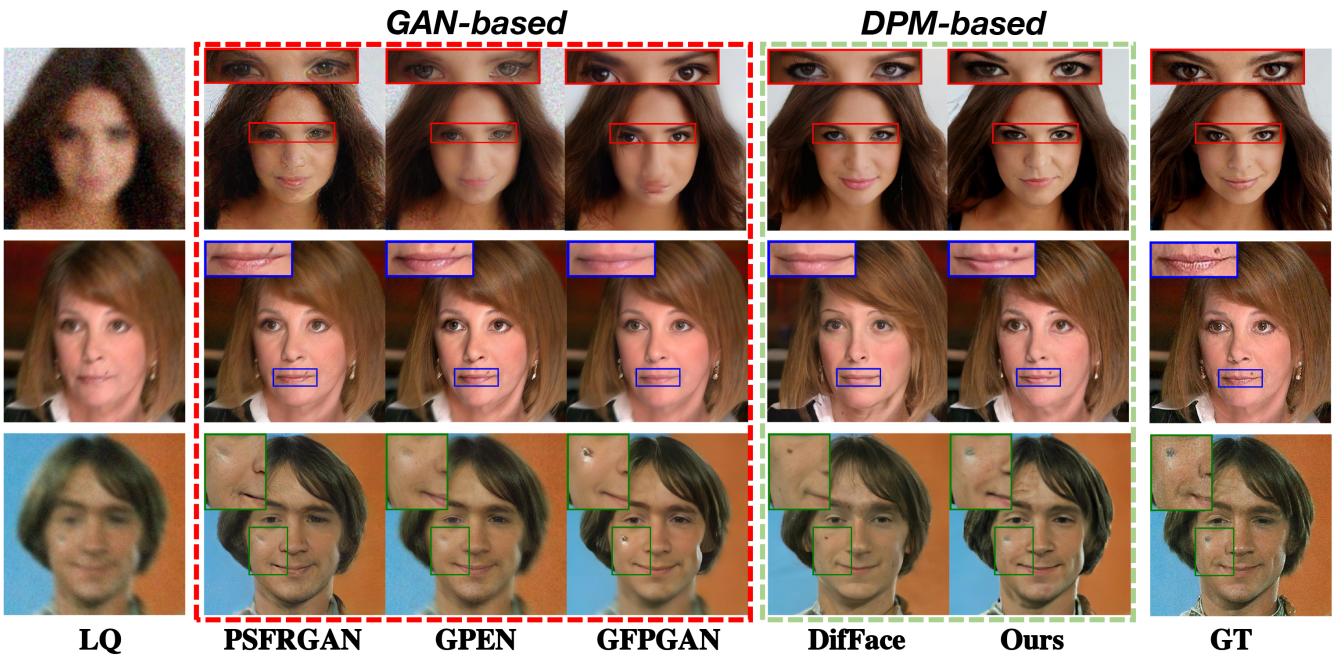

Figure 1: Comparisons of the proposed DiffBFR with state-of-the-art for blind face restoration methods. Left is for low-quality (LQ) images, Middle in the red rectangle for GAN-based results [2, 29, 31], Middle in the green rectangle for DPM-based results [33], and Right for GroundTruth (GT). We can see DiffBFR achieves better restoration details while maintaining the source identity. Better see in color with 2x zoom.

ABSTRACT

Blind face restoration (BFR) is important while challenging. Prior works prefer to exploit GAN-based frameworks to tackle this task due to the balance of quality and efficiency. However, these methods suffer from poor stability and adaptability to long-tail distribution, failing to simultaneously retain source identity and restore detail. In this paper, we propose to introduce Diffusion Probabilistic Model (DPM) for BFR to tackle the above problem, given its superiority over GAN in aspects of avoiding training collapse and generating long-tail distribution. We name the proposed framework as ***DiffBFR***. In particular, DiffBFR utilizes a two-step design, that first restores identity information from low-quality images and then enhances texture details according to the distribution of real faces. This design is implemented with two key components: 1) Identity Restoration Module (**IRM**) for preserving the face details in results. Instead of denoising from pure Gaussian random distribution with LQ images as the condition during the reverse process, we propose a novel truncated sampling method which starts from LQ images

*Corresponding author

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611731

with part noise added. We theoretically prove that this change shrinks the evidence lower bound of DPM and then restores more original details. With theoretical proof, two cascade conditional DPMs with different input sizes are introduced to strengthen this sampling effect and reduce training difficulty in the high-resolution image generated directly. 2) Texture Enhancement Module (**TEM**) for polishing the texture of the image. Here an unconditional DPM, a LQ-free model, is introduced to further force the restorations to appear realistic. We theoretically proved that this unconditional DPM trained on pure HQ images contributes to justifying the correct distribution of inference images output from IRM in pixel-level space. Concretely, truncated sampling with fractional time step is utilized to polish pixel-level textures while preserving identity information. Our experiments demonstrated that the proposed DiffBFR achieves significantly superior results to state-of-the-art methods both quantitatively and qualitatively.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problems**; *Computer vision tasks.*

KEYWORDS

blind face restoration, diffusion probabilistic models

ACM Reference Format:
Xinmin Qiu, Congying Han, Zicheng Zhang, Bonan Li, Tiande Guo, and Xuecheng Nie. 2023. DiffBFR: Bootstrapping Diffusion Model for Blind Face Restoration. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3611731

1 INTRODUCTION

Blind face restoration (BFR) [6, 29, 31] aims to recover high-quality (HQ) face images from the low-quality (LQ) ones. It is an important task in Computer Vision and Graphics communities and is widely applied in various scenarios, *e.g.*, monitoring image restoration, old photos restoration and face image super-resolution, etc. However, this task is very challenging due to the non-deterministic degradation that harms the image quality, such as blurring, noising, down-sampling and compression artifacts.

Previous works for BFR typically rely on Generative Adversarial Networks (GAN). They mainly focus on designing various face-specific priors to tackle the problem, including generative priors [29, 31], reference priors [15, 16] and geometric priors [2, 32]. Although achieving state-of-the-arts, these methods still encounter difficulties in restoring fine-grained facial details while achieving realistic texture, as illustrated in Figure. 1 and Figure. 4. This can be ascribed to the fact that these methods generally need to project degraded images into the latent space of a pre-trained GAN, *e.g.*, StyleGAN [12], while the limited capacity of GANs makes the projection difficult to exactly retain the content details of given images. In addition, GAN-based models face the problem of "training collapse" as the objective function is a min-max function and optimization is difficult. Moreover, due to the poor adaptability of GAN-based methods to long-tail distributed dataset [37], restored faces derived by GAN are prone to change person identities, and they are hard to achieve the balance between image restoration quality and character fidelity maintenance. We conduct experiments on a toy long-tail distribution MNIST to verify this problem and results are shown in Figure. 2, which proves that the GAN-based model [5] fails to cover low-density regions and can not generate local details as the tail feature in the face dataset. It is critical to address these issues for advancing practical applications of BFR in the real world.

To achieve the above goal, we propose to bootstrap diffusion models for blind face restoration in this paper, further pushing forward the frontier of this task. Our main motivation is the superiority of diffusion models over GANs in aspects of avoiding training collapse and generating long-tail distribution. We name our method as ***DiffBFR***. In particular, DiffBFR exploits Diffusion Probabilistic Models (DPM) for enhancing the face-specific prior, considering its great power to produce HQ images in the wild range of distribution. For deriving accurate restoration of LQ faces, DiffBFR utilizes a novel component capturing and texture polishing strategy. Specifically, for component capturing, DiffBRF proposes to denoise from the LQ image and a diffused version of it, which shrinks the evidence lower bound of DPM with theoretical proof and further helps to maintain more details. For texture polishing, DiffBFR relies on the analysis of the similarity of noise space and then exploits rich priors from pure HQ images, which helps to synthesize factual images with natural texture. In this way, DiffBFR completes blind face restoration in two steps: first to restore the content information from LQ images and then to enhance the texture of images, thus producing reliable restoration results.

In particular, DiffBFR is composed of two core modules: (1) Identity Restoration Module (IRM). IRM aims to capture facial information in LQ images. Here, IRM begins with a conditional DPM at low resolution, followed by one conditional super-resolution DPM that upsamples the image. Compared with direct training on large-resolution images, the model can converge faster and obtain better results. During the sampling phase, IRM performs a full reverse diffusion process with low-resolution DPM. For super-resolution DPM, IRM proposes a novel truncated sampling strategy, that is, denoising from intermediate diffused variables, to efficiently preserve more details in results. (2) Texture Enhancement Module (TEM). TEM is designed to polish the texture of images. Specifically, we train an unconditional DPM with pure HQ images and perform denoising starting from the noisy version of the output from IRM. As a result, texture information with a high degree of naturalness is recorded without any impact from LQ images. The use of TEM sharpens the edge structure and further forces the restorations to appear realistic.

From the perspective of both experimental exploration and theoretical derivation, we show that the proposed DiffBRF effectively deploys diffusion models for solving the blind face restoration problem, which not only reduces the training difficulty and training time of the whole model, but also provides less degradation serious conditional input for Truncated Sampling Module. Our contributions can be summarized into three folds:

(1) To the best of our knowledge, we are the first to propose the application of pure diffusion models to the task of blind face restoration, motivated by its superiority over GANs on avoiding training collapse and generating long-tail distribution.

(2) We present two novel modules in DiffBFR: Identity Restoration

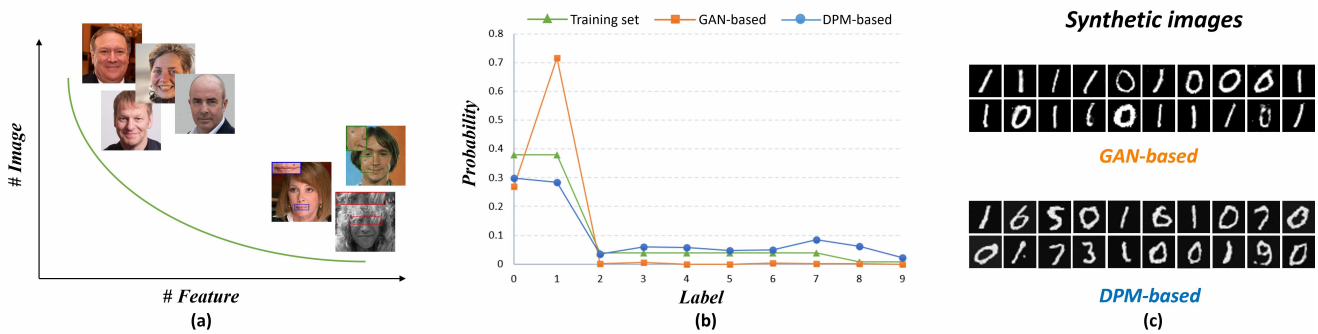

Figure 2: Illustration of long-tail challenge in the BFR task and motivation for our solution based on DPM. (a) The faces sampled from the low-density tail regions in the BFR dataset often comprise novel features, *e.g.*, moles or long fringe, which are hard for existing methods. (b) To address the long-tail challenge, we first investigate the capacities of frequent generative models on a toy long-tail MNIST dataset with 28×28 resolution, where images with labels other than 0 and 1 are partially dropped. (c) The random syntheses combined with statistical data in (b) demonstrate that the GAN-based model fails to synthesize high-fidelity datapoints from low-density regions , while the DPM-based model shows promising results in addressing this problem.

Module (IRM) and Texture Enhancement Module (TEM), which effectively restores high-fidelity facial details while maintaining person identities. Additionally, we also theoretically proved that they can yield better recovery results in the inference process.
(3) Through extensive experiments, we demonstrate that DiffBFR sets new state-of-the-arts on multiple benchmarks for the blind face restoration task.

2 RELATED WORK

Image Restoration. Image Restoration usually includes super-resolution [38], denosing [34], deblurring [13], compression removal [4] and their random combination and so on, which is classical research in the field of computer vision. In the past, most image restoration problems were based on the image degradation model known to give corresponding restoration methods, such as DnCNNs [34], DeblurGAN [13], etc. However, in the real world, the degradation causes of LQ images that need to be restored are mostly unknown. How to restore images whose degradation ways are unknown is an important challenge in this research field in recent years.

Blind Face Restoration. Blind Face Restoration [6, 29, 31] is an important branch in the field of Image Restoration. Its task objective is to restore low-quality (LQ) face images into high-quality (HQ) ones on the premise that degradation models and parameters are completely unknown. In recent years, great breakthroughs have been made in the BFR task, such as the method based on geometric prior of face [2, 32], the method based on reference prior [15, 16], and so on. GFPGAN [29] and GPEN [31] embed face prior information using a GAN-based generation model which uses an encoding-decoding frame. PSFRGAN [2] combined the structural features of face segmentation and proposed a GAN-based progressive restoration network. VQFR [6] combines the classical dictionary-based method with the recent vector quantization (VQ) technology.

Diffusion Probability Models. In the past few years, GAN-based generative models have been almost the mainstream, and after the proposal of Denoising Diffusion Probabilistic Models (DDPM) [9, 20] and Denoising Diffusion Implicit Models (DDIM) [26], the generative model based on diffusion models [25] has become a breakthrough in the field of computer vision with its excellent image generation quality advantage [14, 22, 39]. GAN-based model training is prone to collapse, which is avoided by the diffusion model method. This diffusion-based approach has attracted considerable attention from computer vision and natural language processing to graphic analysis. CARD [7] proposed a classification and regression diffusion model, combining a conditional generation model based on denoising diffusion and a pre-trained conditional mean estimator to predict the data distribution under a given condition. Inspired by CLIP [23], GLIDE [21] explored real image synthesis with text conditions and found that diffusion models with class-free guidance produced high-quality (HQ) images that included a wide range of learned knowledge. With the help of a variational auto-encoder framework, the diffusion model of latent space training is established by LSGM [28]. SegDiff [1] extends the diffusion model to perform image-level segmentation by summarizing feature maps from the diffusion probabilistic encoder and the image feature encoder.

3 PRELIMINARIES

In this section, we briefly introduce fundamental notations and definitions to facilitate comprehension [9, 10, 24] of our proposal.

Denoising Diffusion Probability Models. DDPM [9] establishes a relationship between a complex distribution $p(y)$ and the Gaussian distribution $N(0, I)$ using forward and reverse Markov chains. Following the convention, we denote y as y_0, and the forward process generates latent variables $y_1, \ldots, y_T$ through

$$q(y_t|y_{t-1}) = N(y_t; \sqrt{\alpha_t}y_{t-1}, (1-\alpha_t)I). \tag{1}$$

where $\{\alpha_t\}_{t=1}^T$ is a fixed variance schedule rather than learned parameters. The forward process holds the property

$$q(y_t|y_0) = N(y_t; \sqrt{\gamma_t}y_0, (1-\gamma_t)I), \tag{2}$$

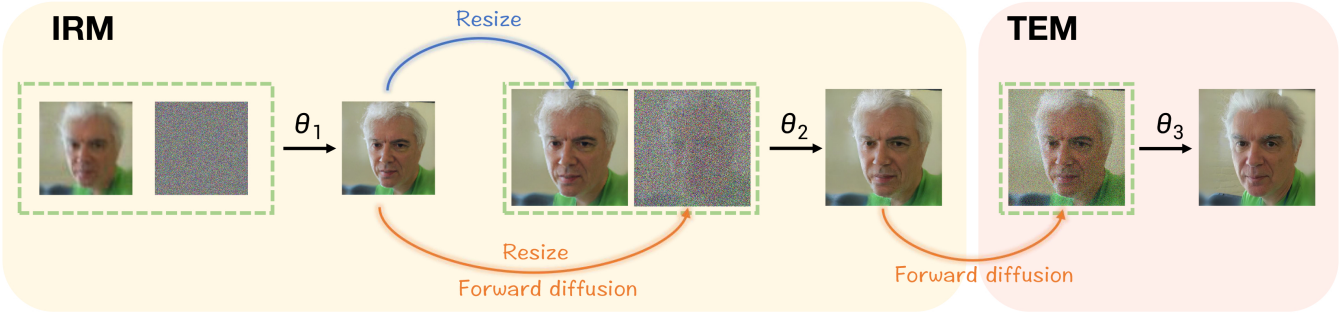

Figure 3: Sampling process of the proposed *DiffBFR* for blind face restoration task. In essence, DiffBFR is a cascaded diffusion model: Given a LQ face, an Identity Restoration Module (*IRM*) enriches the facial details at both low- and high-resolution successively, and a Texture Enhancement Module (*TEM*) further polishes the realistic texture of the image to predict the HQ face. The DPM-based design of DiffBFR confers advance in performance verified by both theoretical and practical evidence.

where $\gamma_t = \prod_{i=1}^{t} \alpha_i$. The reverse process starts from y_T to sample the real data y_0 sequentially through

$$p_\theta(y_{t-1}|y_t) = N(y_{t-1}; \mu_\theta(y_t, t), \Sigma_\theta(y_t, t)), \tag{3}$$

where μ_θ is a parameterized function to be trained for maximizing evidence lower bound (ELBO) of $p(y_0)$, and $\Sigma_\theta = \sigma_t^2 I$ where σ_t is usually a pre-defined constant related to the variance schedule. Further, by decomposing μ_θ into a linear combination of x_t and the noise approximator ϵ_θ, the generative process can be expressed in another form:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}}\epsilon_\theta(x_t, t)) + \sigma_t \epsilon, \tag{4}$$

where $\epsilon \sim N(0, I)$, which suggests that each generation step is stochastic. Similarly, a conditional distribution $p(y_0|x)$ can be approximated by the diffusion process:

$$p_\theta(y_{t-1}|y_t, x) = N(y_{t-1}; \mu_\theta(y_t, t, x), \Sigma_\theta(y_t, t)). \tag{5}$$

Cascaded Diffusion Model. Cascaded diffusion model [10] (CDM) is an effective method to scale a diffusion model to high-dimension distribution. Specifically, for a high-resolution image y_0, an extra latent variable (*e.g.*, down-sampled image) z_0 that is easier to learn than y_0 is introduced, thus we can reformulate the generative process of y_0 as

$$p(y_0) = \int p(y_0|z_0)p(z_0)dz_0, \tag{6}$$

which corresponds to a two-stage cascaded model. In this way, CDM first learns a diffusion model as Eq.(3) for low-resolution image z_0, then learns a conditional diffusion model as Eq.(4) to sample y_0 from $p(y_0|z_0)$. In practice, the cascaded process can be divided into multiple stages by inserting additional latent variables.

4 METHODOLOGY

In this section, we present DiffBFR, a diffusion probability model designed to address the BFR task. As depicted in Figure. 3, DiffBFR primarily comprises two fundamental modules: the Identity Restoration Module (IRM) and the Texture Enhancement Module (TEM). IRM learns to straightforwardly enhance facial identity details at both low- and high-resolution levels, achieving superior identity preservation. TEM further refines the realistic texture of the image with a DPM-based facial texture prior, enabling the prediction of HQ face images.

Unlike previous methods that project LQ images into the vectorized and compact latent space of pre-trained GANs, potentially resulting in texture and identity information loss, the proposed DiffBFR offers a more intuitive solution for enhancing image details in a non-compressed and expressive latent space, while preserving facial details. In the following, we begin with a comprehensive analysis of the BFR task and DiffBFR's mechanism (Sec. 4.1), then delve into the technical details of the IRM (Sec. 4.3) and TEM (Sec. 4.4) components to illustrate the advantages of our proposal.

4.1 Long-tail challenge in BFR task

We approach the BFR problem from the lens of conditional generation: Given a dataset consisting of various LQ-HQ image pairs, we aim to learn a conditional distribution $p(y_0|x_0)$, in which x_0 and y_0 denote the LQ and HQ variables, respectively. Empirically, the data in BFR dataset are typically scattered across a high-dimensional space with a long-tail distribution [37]: The head region of distribution only comprises a limited number of normal cases, whereas the long-tail region consists of numerous hard cases, *e.g.*, grayscale image and face with moles. Unlike in the classification task [27], the low-level feature appearing on the tail part refers to attributes that less influence the identity, but is important for visual effects.

We state that *learning such a long-tail distribution $p(y_0|x_0)$ poses a significant challenge for existing BFR methods.* As evidenced by Figure. 1, previous GAN-based works cannot well tackle the samples residing in both head and long-tail regions, resulting in obvious over-smoothing texture as well as distorted content compared to the GroundTruth. DiffFace [33], a concurrent DPM-based method, also encounters similar issues. Given the practical limitations of existing methods, it is significant to tackle the challenge of advancing the frontier of BFR.

4.2 DiffBFR: Diffusion model for BFR task

In this part, we explore overcoming the challenge via a reasonable design to well approximate the long-tail distribution $p(y_0|x_0)$. Our proposal named DiffBFR, a DPM-based model for BFR task, has two main advantages: **(i)** Clear theoretical strengths and interpretability.

(**ii**) Concise and easy to train in practice. As presently there are two mainstream generative models to learn a distribution, *i.e.*, GAN and DPM, we first answer the following question to strengthen the rationale of choosing DPM rather than GAN as the base model:

DPM or GAN, which one is the most promising to solve the long-tail challenge? As shown in Figure. 2, we provide a toy dataset and models to explain that DPM would be the solution. At first, we construct a long-tail MNIST dataset with 28×28 resolution. Compared with the vanilla MNIST [3], we partially discard some samples, such that the images with labels 0 and 1 have a higher density and the others have a lower density. Then, we train toy DDPM [9] and GAN [5] on the long-tail dataset with 28×28 resolution, in which generators have similar numbers (~1.5 million) of parameters. After that, we count the labels of random samples from trained DPM and GAN. The results demonstrate that the DPM is promising to align the long-tail distribution reasonably, whereas GAN tends to fit the head region with high density, resulting in a very low probability of label generation in the long-tail, and even a few categories are barely generated anymore. Therefore, we design DiffBFR as a DPM-based model to better solve the challenge.

Cascaded structure in DiffBFR. Although DDPM performs better in the toy long-tail dataset, in practice the large size ($\geq 512^2$ pixels) and scale (>50k) of BFR datasets make it non-trivial to directly apply it to the BFR task. We find that a proper design of cascaded structure can not only enhance training stability [10], but also improve the quality of restoration. Specifically, DiffBFR is based on the reformulation:

$$p(y_0|x_0) = \int p(y_0', x_0'|x_0)p(y_0|y_0', x_0', x_0)dx_0'dy_0', \tag{7a}$$

$$\approx \int \underbrace{p(x_0'|x_0)p(y_0'|x_0')}_{IRM}\underbrace{p(y_0|y_0')}_{TEM}dx_0'dy_0'. \tag{7b}$$

Herein, we introduce two new intermediate variables x_0' and y_0' with the same shape of x_0 and y_0, respectively. The formulation of DiffBFR first follows and inherits the advantages of CDM in training speed and stability, where each conditional and unconditional distribution in Eq.(7b) can be approximated by SR3 [24] or DDPM [9]. Moreover, beyond just outperforming in model training, we note that each module in DiffBFR with its specific design will enhance the prediction for the BFR task. The first one is called the Identity Restoration Module where (IRM) upsamples the LQ image x_0 to gradually arrive at the resolution of y_0 while enriching the facial details, and the second one called TEM exploits the diffusion-based facial prior to further refining the texture details. Both IRM and TEM are equipped with *truncated sampling strategies*, alleviating the unfaithful results due to excessive noise in Eq.(2). The remaining part elaborates on technical details.

4.3 Identity Restoration Module

Given each training LQ-HQ pair (x_0, y_0), the IRM learns the cascaded conditional distribution to map LQ image x_0 into the high-resolution image with two steps. The first stage first enriches the facial details at a low resolution as same as x_0, where a DDPM is trained with the objective

$$\min_{\theta_1} \mathbb{E}_{x_0,\epsilon\sim N(0,I),t\sim\text{Uniform(1, T)}} \left\|\epsilon - \epsilon_{\theta_1}(x_t', t, x_0)\right\|_2^2. \tag{8}$$

x_0' is the low-resolution GroundTruth downsampled from y_0 with a scale factor r, *i.e.*, $x_0' = [y_0]\downarrow_r$, and x_t' is the noisy image of x_0' sampled from Eq.(2). We denote the sample from learned distribution as $\tilde{x}_0'$. Then a DDPM is trained with the following objective

$$\min_{\theta_2} \mathbb{E}_{x_0,\epsilon\sim N(0,I),t\sim\text{Uniform(1, T)}} \left\|\epsilon - \epsilon_{\theta_2}(y_t', t, \tilde{x}_0')\right\|_2^2. \tag{9}$$

We provide more training details of ϵ_{θ_1} and ϵ_{θ_2} in Experiments (Sec. 5) and Supplementary Materials.

Truncated sampling. The sampling strategy in the reverse process [17] based on Eq.(4) has a crucial impact on the quality of results. For the BFR task, we find the way starting with $y_T' \sim N(0, I)$ to sample from $p_\theta(y_{t-1}'|y_t', \tilde{x}_0')$ subsequently cannot exploit the full potential of the trained DDPMs, where the final result y_0' are probably unfaithful to $\tilde{x}_0'$ in terms of identity. Therefore, we propose a truncated sampling strategy in the conditional frame to improve it. The reverse process will be conditioned on y_{N_1}, where the truncated time $N_1 < T$. In the following proposition, we provide a theoretical analysis of the advantage of truncated sampling compared with vanilla sampling.

Proposition 1. *Given a LQ image x_0 and HQ image y_0, we denote the evidence lower bound (ELBO) of vanilla diffusion, and diffusion with truncated sampling as L_{DDPM} and L_{IRM}, respectively. Then, we have*

$$L_{DDPM} \leq L_{IRM}. \tag{10}$$

Proposition 1 shows that for conditional DDPM, the change of truncated sampling can shrink the ELBO of the model. Furthermore, it can be proved that the higher the quality of the condition input $\tilde{x}_0'$, the closer it is to y_0, the more accurate the restored image will be. This explains why we need to restore low-resolution images first in IRM. In a nutshell, we design IRM as follows: the restoration preprocess on low-resolution images provides an input, so that the conditional DPM on high-resolution ones can generate higher-quality images with these effective sampling changes.

4.4 Texture Enhancement Module

Despite the delicate facial details can be well restored via IRM, we experimentally find that the results usually retina some weird texture, such as the edge on the corners of the eyes, teeth and other facial features, which are obvious to impede the visual effect. We conjecture that this unnatural texture may result from the excessive restoration of IRM. In the end, we find imposing a diffuse-based facial prior to restored faces from IRM can greatly remove texture weakness. We train an *unconditional* DDPM with the objective

$$\min_{\theta_3} \mathbb{E}_{y_0,\epsilon\sim N(0,I),t\sim\text{Uniform(1, T)}} \left\|\epsilon - \epsilon_{\theta_3}(y_t, t)\right\|_2^2. \tag{11}$$

In this way, the sampling starts from $y_{N_2} \sim q(y_{N_2}|y_0')$ that sampled from Eq.(2) indeed formulate $p(y_0|y_0')$ to enhance the texture details of restored faces, which names TEM.

Moreover, by cooperating with Fréchet Inception Distance in theory, we prove that TEM can effectively correct the distribution of the restoration images.

Figure 4: Qualitative comparisons on the CelebA-Test for blind face restoration and from left to right: low-quality image, PULSE [18], PSFRGAN [2], GPEN [31], GFPGAN [29], VQFR [6], DifFace [33], our DiffBFR and GroundTruth. Our DiffBFR performs well in both detail complement and hue preservation. Zoom in for best view.

Table 1: Quatitative comparison on CelebA-Test with 3000 images randomly for blind face restoration. Red, underline and blue indicate the best, the second best and the third best performance.

Metrics	Input(LQ)	Methods							
		DFDNet[15]	PULSE[18]	GPEN[31]	GFPGAN[29]	PSFRGAN[2]	VQFR[6]	DifFace[33]	**DiffBFR**(ours)
SSIM↑	0.6460	0.6444	0.6102	0.6777	0.6827	0.6213	0.6382	0.6494	0.6553
PSNR↑	24.921	23.300	21.619	25.423	25.401	24.596	23.568	24.055	24.748
FID↓	93.564	39.649	45.940	22.507	20.676	26.050	17.862	19.653	16.490
NIQE↓	9.1407	6.4226	7.3754	6.7775	6.7324	5.6114	5.9606	6.1638	5.5990
LPIPS↓	0.5953	0.3901	0.4209	0.2956	0.2823	0.3101	0.2616	0.3052	0.2535

Proposition 2. *Assume that the LQ image input is x, the HQ image is y, and the inference image is y'. It can be proved that the FID of the resulting image distribution after TEM is lower than that before TEM. We have*

$$FID(x, y) > FID(y', y). \tag{12}$$

Proposition 2 is precisely proving that the FID of the inference that images distribution after TEM is lower than that before TEM, and the obtained inference images have a more similar distribution than HQ images on the whole.

5 EXPERIMENTS

In this section, we introduce the training dataset, testing dataset in Sec. 5.1 and specific experimental results comparison in Sec. 5.2. We perform ablation studies to demonstrate the effectiveness of the proposed IRM and TEM in Sec. 5.3.

5.1 Datasets

Training Datasets. We choose FFHQ [12] as the training dataset, which contains 70,000 high-quality PNG format face images with 1024×1024 resolution. In this experiment, we resize all images to 512×512 to train face restoration at this resolution.

Since our DiffBFR is supervised training, the corresponding LQ-HQ image pairs are required. We use generated random degradation model to simulate LQ images in the real world. Its generation formula [29, 35] is shown in Eq.(13), where y is the HQ image, k_σ is the Gaussian blur kernel, r represents the down-sampling scale factor, and q represents the JPEG compression of the image with quality factor q. In order to keep the experimental results directly comparable, the parameters σ, r, δ, q are randomly sampled from {0.1: 10}, {0.8: 8}, {0: 20}, {60: 100}, respectively, to align with the experimental environment of recent methods for BFR task. We also

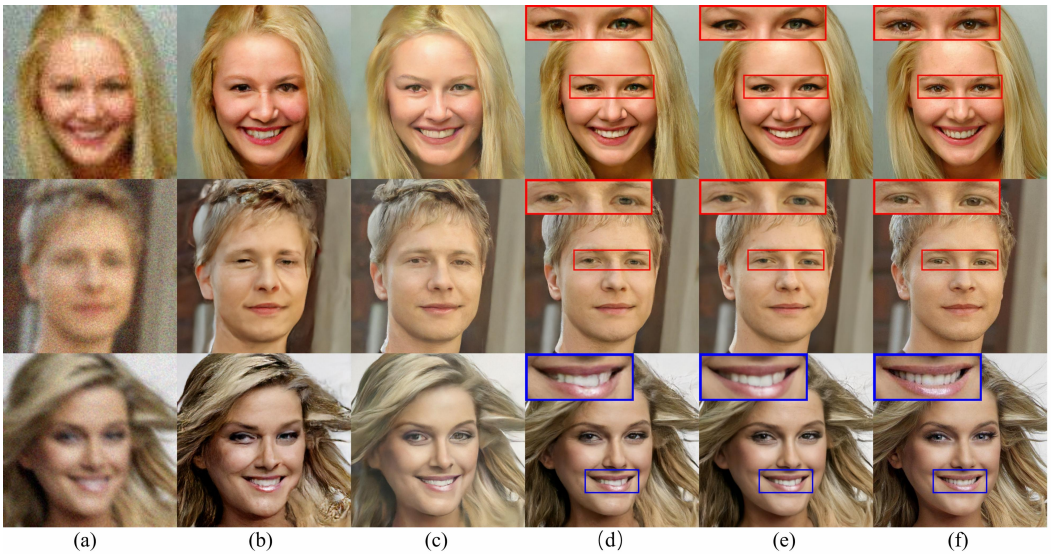

Figure 5: Qualitative comparisons on CelebA-Test for blind face restoration in ablation results. (a) LQ images, (b) IRM-s: 1-stage DPM without cascade, (c) IRM-c: 2-stage DPM with cascade, (d) IRM-c-t: 2-stage DPM which is added truncated sampling module in the second stage, namely IRM(-2), (e) TEM-w: 3-stage DPM which is added unconditional justify module in TEM, (f) GT images. Zoom in for best view.

add gray color probability during the training process for color adaptation and augment data with the horizontal flip.

$$x = [(y \otimes k_\sigma) \downarrow_r + n_\delta]_{\text{JPEG}_q} \tag{13}$$

Testing Datasets. We choose CelebA-Test as the testing dataset, which contains 3,000 HQ images randomly sampled from CelebA-HQ [11] with the resolution of 512×512. Similarly, the corresponding random LQ images are generated for evaluation by using the degradation model in Eq.(13) and the same set of parameters used in the training dataset. Our method and other state-of-the-art methods are tested on the same CelebA-Test dataset to observe their quantitative comparisons and qualitative comparisons.

5.2 Comparisons with State-of-the-art Methods

Comparison Methods. During the experiments, we noted the concurrent work, DifFace [33], which is also included in the comparisons. We compare DiffBFR with seven recent BFR methods, including DFDNet [15], PULSE [18], PSFRGAN [2], GFPGAN [29], GPEN [31], VQFR [6], and DifFace.

Metrics. We quantitatively compare the differences between our method and state-of-the-art methods using five widely-used metrics, including SSIM [30], PSNR, FID [8], NIQE [19], and LPIPS [36]. Among them, NIQE is a no-reference metric. SSIM and PSNR are pixel-wise similarity measures, while FID, NIQE and LPIPS are perceptual measures.

Quantitative Results. As shown in Table. 1, the comparison results on the CelebA-Test are summarized and our method shows better results in quantitative results. DiffBFR achieves the best FID, NIQE and LPIPS scores, indicating that our restoring results are close to the real face image distribution and the natural image distribution and maintain the perceptual approximation to GroundTruth. However, the pixel-wise metrics SSIM and PSNR are not highly correlated with the subjective evaluation of human observation. DiffBFR only maintains a relatively similar degree with recent state-of-the-art methods in these two metrics to achieve the basic goal of the restoration task, which is not good at these two measures.

Table 2: Ablation study results on CelebA-Test for blind face restoration. *IRM-s*: use 1-stage DPM in IRM; *IRM-c*: use 2-stage cascade DPM in IRM; *IRM-c-t*: change the sampling process in the second stage in truncated sampling; *TEM-w*: add the advanced unconditional DDPM in TEM.

Method	SSIM↑	PSNR↑	FID↓	NIQE↓	LPIPS↓
IRM-s	0.5266	22.8438	31.3126	6.3403	0.4873
IRM-c	0.5879	21.5117	24.2364	5.8538	0.3378
IRM-c-t	0.6494	24.727	19.6023	5.4831	0.2546
TEM-w	0.6553	24.7485	16.4902	5.5990	0.2535

Qualitative Results. Figure. 4 shows the restoration effect comparison of color images and gray images. Obviously, our method can see the restoration ability of the face in the visual image. Due to the inclusion of the conditional module, DiffBFR maintains quite good results in fidelity. From the figure, we can see that in LQ images with serious degradation, DiffBFR is able to obtain inference images without blurring and significant noise residual. Additionally, for color images and gray images, DiffBFR can maintain the same color intensity as the GroundTruth as much as possible, which plays an important role in the restoration of light and shadow effects in image restoration. From Figure. 4, we can see that PULSE [18] changes the identity during the restoration process, and the restored face of the severely degraded image is not the same person from the human point of view. DFDNet [15] has a limited ability to restore the face structure, and many details keep the blurred part in the LQ image, which can not supplement the clearer HQ image. PSFRGAN [2], GPEN [31] and GFPGAN [29] are all proposed GAN-based methods. It can be seen that their restoration is more in line with

the view of the real world in terms of the realness of the face than traditional methods, but it is not as good as the method based on diffusion models (namely DifFace and our method) in maintaining and predicting the original image information.

5.3 Ablation Studies

To better understand the roles of different components of DiffBFR, we conduct ablation studies. The first part is denoted by IRM-s, which used 1-stage DPM without introducing a cascade approach. The second part is denoted by IRM-c, which used 2-stage CDM with the traditional sampling process. The third part is denoted by IRM-c-t, which used 2-stage CDM with the Truncated Sampling Module in the second stage, that is the complete IRM in our DiffBFR. The last part is denoted by TEM-w, which added the advanced unconditional DDPM in TEM as the justify module.

We perform BFR on the CelebA-Test dataset to evaluate different components of DiffBFR. The LQ images are synthesized by the degradation model in Eq.(13). As shown in Figure. 5, IRM-s does not apply to the degradation model with uncertain parameters and combining multiple degradation modes, and the obtained inference image still has residual blur and noise, and the improvement of image resolution is not obvious. IRM-c decomposes the restoration process in different resolutions, and it can be clearly seen from the image that the blur degree is reduced, but there is still obvious noise residual. To remove the noise residue in the image and generate relatively detailed face information faithfully, IRM-c-t changes its sampling process. It can be clearly seen from the output of IRM-c-t that the noise added in the diffusion process is easy to be left when restoring the severely degraded image. Table. 2 lists metric results of ablation experiments. We found that after adding Truncated Sampling Module in IRM, the image noise is effectively reduced from the qualitative perspective, and FID and LPIPS are significantly reduced from the quantitative perspective. TEM-w achieves considerable results as shown in the Table. 2, reducing indicators FID and LPIPS effectively and making the image distribution close to the real face image distribution. In Figure. 5, it is shown that this component restores local over-smoothness in details such as eyes and teeth, and the detail contour of the face is more natural and in line with the real face. Overall, DiffBFR shows superior performance to these partial components, demonstrating the efficacy of our theoretical proof.

Additionally, we assume our DiffBFR three stages respectively to explore extra parameters. In the sampling process of IRM which contains two stages, low-resolution in IRM(-1) and high-resolution in IRM(-2), the selection of the super-parameter N_1 depends on the output quality of IRM(-1) and the precision of network prediction in IRM(-2). The ablation results of the value of N_1 and N_2 are shown in Table. 3 and Table. 4.

5.4 Discussion

Advantages. **(1)** Our method DiffBFR is closer to GroundTruth in the restoration effect, especially in the image color intensity and light intensity, which restores the original image to a greater extent. **(2)** Inference images of DiffBFR are more realistic than those of GAN-based methods. Restored images based on GAN methods pay attention to the integrity of prior knowledge, which is easy to cause huge changes to the whole facial features, while our method restores the details and retains the structural information of the original HQ image simultaneously. **(3)** One low-quality image can directly and reasonably correspond to several different HQ images, so the fixed mapping relationship limits the various possibilities of restoration. While DiffBFR has a certain randomness in the sampling process, which can give multiple reasonable reasoning images at the same time to deal with various possible restoration scenarios.

Table 3: Ablation study results about N_1 in the sampling process of IRM(-2) on CelebA-Test. We choose $N_1 = 1000$.

N_1	SSIM↑	PSNR↑	FID↓	NIQE↓	LPIPS↓
200	0.6759	25.4322	22.5438	5.6177	0.2759
600	0.6604	25.0672	18.5485	5.5000	0.2617
1000	0.6494	24.727	19.6023	5.4831	0.2546
1400	0.6395	24.2919	21.4712	5.4669	0.2725
1800	0.62746	23.2766	22.8638	5.6274	0.2880

Table 4: Ablation study results about N_2 in the sampling process of TEM on CelebA-Test. We choose $N_2 = 100$.

N_2	SSIM↑	PSNR↑	FID↓	NIQE↓	LPIPS↓
80	0.6555	24.7671	16.4227	5.5868	0.2534
100	0.6553	24.7485	16.4902	5.5990	0.2535
120	0.6550	24.7236	16.5395	5.5956	0.2540
150	0.6535	24.6658	16.8194	5.6041	0.2551

Limitations. **(1)** Our method inherits the characteristics of diffusion models in the inference process, and runs for a long time. Although the Truncated Module reduces the sampling time by half, it is still longer than the running time of GAN-based methods. It needs to be further optimized for accelerated sampling in the future. **(2)** Compared to SR3 [24], a super-resolution method based on diffusion models, the parameter scale of our training model is larger, which is caused by the cascaded multi-stage model, and also for the task of image restoration with more severe degradation rather than just clean image super-resolution.

6 CONCLUSION

We have proposed DiffBFR, a face image restoration model for blind degradation based on pure diffusion models, motivated by its superiority over GANs on avoiding training collapse and generating long-tail distribution. By embedding prior into diffusion models, our model learned to generate HQ face images from randomly severely degraded ones. Specifically, we proposed two modules IRM and TEM to restore fidelity and realistic details respectively. The derivation of the theoretical boundary and the demonstration of the experimental images show the advantages of the model, and compared with recent SOTA methods, the qualitative and quantitative results are better. In the future, we will extend DiffBFR to much more severe degraded images to restore correct and realistic details.

ACKNOWLEDGMENTS

This paper is supported by the National key research and development program of China (2021YFA1000403), and the National Natural Science Foundation of China (Nos. U19B2040, 11991022).

REFERENCES

[1] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390* (2021).

[2] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. 2021. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*. 11896–11905.

[3] Li Deng. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* (2012).

[4] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2015. Compression artifacts reduction by a deep convolutional network. In *ICCV*. 576–584.

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

[6] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. 2022. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*. Springer, 126–143.

[7] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. 2022. Card: Classification and regression diffusion models. In *NeurIPS*. 18100–18115.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*. 6840–6851.

[10] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research* 23, 1 (2022), 2249–2281.

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*.

[12] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.

[13] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*. 8183–8192.

[14] Bonan Li, Zicheng Zhang, Xuecheng Nie, Congying Han, Yinhan Hu, and Tiande Guo. 2023. StyO: Stylize Your Face in Only One-Shot. *arXiv preprint arXiv:2303.03231* (2023).

[15] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. 2020. Blind Face Restoration via Deep Multi-scale Component Dictionaries. (2020), 399–415.

[16] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. 2018. Learning Warped Guidance for Blind Face Restoration. (2018), 272–289.

[17] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*.

[18] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*. 2437–2445.

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

[20] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*. 8162–8171.

[21] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*. 16784–16804.

[22] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*.

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

[24] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4713–4726.

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

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[27] Yichun Tai, Hailin Shi, Dan Zeng, Hang Du, Yibo Hu, Zicheng Zhang, Zhijiang Zhang, and Tao Mei. 2023. Multi-Agent Semi-Siamese Training for Long-Tail and Shallow Face Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6, Article 196 (jul 2023), 20 pages. https://doi.org/10.1145/3594669

[28] Arash Vahdat, Karsten Kreis, and Jan Kautz. 2021. Score-based generative modeling in latent space. In *NeurIPS*. 11287–11302.

[29] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards Real-World Blind Face Restoration with Generative Facial Prior. (2021), 9168–9178.

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[31] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *CVPR*. 672–681.

[32] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. 2018. Face super-resolution guided by facial component heatmaps. In *ECCV*. 217–233.

[33] Zongsheng Yue and Chen Change Loy. 2022. DifFace: Blind Face Restoration with Diffused Error Contraction. *arXiv preprint arXiv:2212.06512*.

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

[35] Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2018. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*. 3262–3271.

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 586–595.

[37] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2023. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

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

[39] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. 2023. Towards Consistent Video Editing with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2305.17431* (2023).

A PROOFS

Proposition 3 (**IRM**). *Given a LQ image x_0 and HQ image y_0, we denote the evidence lower bound (ELBO) of vanilla diffusion, and diffusion with truncated sampling as L_{DDPM} and L_{IRM}, respectively. Then, we have*

$$L_{DDPM} \leq L_{IRM}. \tag{14}$$

Proof. First, we can give the ELBO expression (15) for conditional DDPM.

$$\begin{aligned} \log p_\theta(y_0|x) &\geq E_{q(y_{1:T}|y_0)}[\log \frac{p_\theta(y_{0:T}|x)}{q(y_{1:T}|y_0)}] \\ &= -D_{KL}(q(y_T|y_0)||p(y_T|x)) \\ &- \sum_{t=2}^{T} D_{KL}(q(y_t|y_{t-1})||p_\theta(y_{t-1}|y_t,x)) \\ &+ \log p_\theta(y_0|y_1,x) = L_{DDPM} \end{aligned} \tag{15}$$

In formula (15), we focus on the first term, i.e

$$D_{KL}(q(y_T|y_0)||p(y_T|x)) := L_T \tag{16}$$

In the DDPM method without Truncated Sampling Module, this term is

$$\begin{aligned} &\rightarrow D_{KL}(q(y_T|y_0)||p(y_T)) := L_{T|DDPM} \\ &p(y_T) = N(y_T|0,I) \end{aligned} \tag{17}$$

After changing the sampling method (IRM), this term is

$$\begin{aligned} &\rightarrow D_{KL}(q(y_T|y_0)||q(y_T|x)) := L_{T|IRM} \\ &q(y_T|x) = N(y_T|\sqrt{\gamma_T}x, (1-\gamma_T)I) \end{aligned} \tag{18}$$

$$L_{T|IRM} = \frac{1}{2}\frac{\gamma_T}{1-\gamma_T} \parallel x - y_0 \parallel^2 \tag{19}$$

When T cannot take positive infinity, and x as the LQ image itself has partial information, we have

$$L_{T|DDPM} \geq L_{T|IRM} \tag{20}$$

Then we prove the formula 14.

Proposition 4 (**TEM**). *Assume that the LQ image input is x, the HQ image is y, and the inference image is y'. It can be proved that the FID of the resulting image distribution after TEM is lower than that before TEM. We have*

$$FID(x,y) > FID(y',y). \tag{21}$$

Proof. Because $x = \Phi(y)$ is a pair of LQ-HQ image pairs, a sufficiently large N can be satisfied by the formula 22 during the diffusion process of adding noise.

$$FID(x,y) > FID(x_N, y_N) \tag{22}$$

For x_N and y_N, in the same unconditional denoise process, we could sample x_0 and y_0, respectively. In addition, since the unconditional diffusion model maps the completely Gaussian random distribution to the real distribution of data in the sampling process, namely the HQ image distribution here, we can obtain

$$|FID(x_0,y_0) - FID(x_N,y_N)| < |FID(x,y) - FID(x_N,y_N)| \tag{23}$$

At this time, $x_0$is inference image and $y_0$is HQ image, and the formulas 22 and 23 can be deduced

$$FID(x,y) > FID(x_0,y_0) = FID(y',y). \tag{24}$$

Algorithm 1: Inference process

Input: Low-quality image x, prediction networks $\epsilon_{\theta_1}(x'_t, t, x_0)$, $\epsilon_{\theta_2}(y'_t, t, resize(x'_0))$ and $\epsilon_{\theta_3}(y_t, t)$, parameter N_1, N_2
Output: inference high quality image y_0

$x'_T \sim N(0,I)$
for $t = T, ..., 1$ **do**
  $\epsilon \sim N(0,I)$ if $t > 1$, else $\epsilon = 0$
  $x'_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x'_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}}\epsilon_{\theta_1}(x'_t, t, x_0)) + \sqrt{1-\alpha_t}\epsilon$
end
$y'_{N_1} \sim q(y'_{N_1}|y'_0 = resize(x'_0))$
for $t = N_1, ..., 1$ **do**
  $\epsilon \sim N(0,I)$ if $t > 1$, else $\epsilon = 0$
  $y'_{t-1} = \frac{1}{\sqrt{\alpha_t}}(y'_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}}\epsilon_{\theta_2}(y'_t, t, resize(x'_0)) + \sqrt{1-\alpha_t}\epsilon$
end
$y_{N_2} \sim q(y_{N_2}|y_0 = y'_0)$
for $t = N_2, ..., 1$ **do**
  $\epsilon \sim N(0,I)$ if $t > 1$, else $\epsilon = 0$
  $y_{t-1} = \frac{1}{\sqrt{\alpha_t}}(y_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}}\epsilon_{\theta_3}(y_t, t)) + \sqrt{1-\alpha_t}\epsilon$
end

B INFERENCE PROCESS

Algorithm 1 describes the inference process of DiffBFR, corresponding to Figure. 3.

C ADDITIONAL RESULTS ON MNIST

We present additional experimental details and results on the toy MNIST dataset, referring to Section 4.2, the main paper of our study. Table. 5 shows parameter details of two toy models.

Table 5: Parameter details of toy models used in the main paper.

Method	GAN-based	DPM-based
Total parameters	1.51M	1.44M
Total memory	1.44Mib	1.38Mib

D ADDITIONAL COMPARISON RESULTS

Table. 6 provides more details of models used in Section 5 of the main paper. In Figure. 6, we present additional comparison results of DiffBFR against other state-of-the-art methods including PULSE, DFDNet, PSFRGAN, GPEN, GFPGAN, VQFR, and DifFace.

Table 6: Configuration details in Section 5 of the main paper. Both Model-1 and Model-2 are trained on NVIDIA RTX 3090.

Details		Model-1/IRM(-1)	Model-2/IRM(-2)	Model-3/TEM
diffusion model	Input size	128×128	512×512	512×512
	Output size	128×128	512×512	512×512
	conditional	true	true	false
	Time step	2000	2000	1000
	Beta schedule	[1e-6, 1e-2]	[1e-6, 1e-2]	[1e-4, 2e-2]
	Loss type	L_1	L_1	L_1
	Sampler time step started	2000	1000	100
UNet	channel multiplier	[1, 2, 4, 8, 8]	[1, 2, 4, 8, 8, 16, 16]	[1, 2, 4, 8, 8, 16, 16]
	In channel	6	6	3
	Out channel	3	3	6
	Inner channel	64	64	32
	Attention resolutions	[16]	[16]	[32, 16, 8]

Figure 6: Qualitative comparisons on the CelebA-Test for blind face restoration. Our DiffBFR performs well in both detail complement and hue preservation. Zoom in for the best view.