Diffuse and Restore: A Region-Adaptive Diffusion Model for Identity-Preserving Blind Face Restoration

Maitreya Suin
smaitre1@jh.edu

Nithin Gopalakrishnan Nair
ngopala2@jhu.edu

Chun Pong Lau
cplau27@cityu.edu.hk

Vishal M. Patel
vpatel36@jhu.edu

Rama Chellappa
rchella4@jhu.edu

Abstract

Blind face restoration (BFR) from severely degraded face images in the wild is a highly ill-posed problem. Due to the complex unknown degradation, existing generative works typically struggle to restore realistic details when the input is of poor quality. Recently, diffusion-based approaches were successfully used for high-quality image synthesis. But, for BFR, maintaining a balance between the fidelity of the restored image and the reconstructed identity information is important. Minor changes in certain facial regions may alter the identity or degrade the perceptual quality. With this observation, we present a conditional diffusion-based framework for BFR. We alleviate the drawbacks of existing diffusion-based approaches and design a region-adaptive strategy. Specifically, we use an identity preserving conditioner network to recover the identity information from the input image as much as possible and use that to guide the reverse diffusion process, specifically for important facial locations that contribute the most to the identity. This leads to a significant improvement in perceptual quality as well as face-recognition scores over existing GAN and diffusion-based restoration models. Our approach achieves superior results to prior art on a range of real and synthetic datasets, particularly for severely degraded face images.

1. Introduction

Degraded face images are frequently encountered in the wild, usually involving a combination of various complex factors such as low resolution, blur, noises, encoding artifacts, etc. Blind face restoration is a highly ill-posed image restoration problem that aims at restoring high-quality face images from low-quality counterparts without knowing the specific degradation [34, 37]. Conventional methods [1, 2, 26] usually depend on the degradation model and handcrafted priors resulting in sub-optimal performance and limited generalization capability while handling a diverse range of real-world face images. Recently, the focus has shifted towards deep learning-based generative methods that usually exploit large-scale datasets and exhibit superior performance. Majority of existing generative frameworks [24, 27, 27] first project the degraded image to a highly-compressed latent space and aim to predict the clean latent embedding. Such approaches have a few significant disadvantages; for example, it is difficult to accurately project a face image with a limited resolution to a lower-dimensional latent space, often losing finer details. Furthermore, the utilization of adversarial loss can introduce optimization instability, mode collapse, and even unwanted artifacts, thereby causing significant distortion. On the other hand, preserving and recovering the underlying identity is crucial for the BFR task. A few existing works [35] typically use an additional loss function (using a pretrained face-recognition network) on the same generative model to retrieve the identity w.r.t. the ground-truth (GT) image. However, we observe that training a single network to simultaneously optimize both perceptual quality and identity preservation poses substantial challenges. The emphasis on identity preservation often comes at the cost of compromising the visual quality of the output, as simply combining two completely different objective functions may not align well for a single network and all the spatial locations of an image.

In our work, we adopt a novel approach involving two distinct networks, each trained with its unique objective, and subsequently merge their outputs using a region-adaptive strategy. To generate visually appealing restored images, we leverage the power of Denoising Diffusion Probabilistic Models (DDPM) [10, 14]. DDPMs have garnered acclaim for their ability to produce high-quality outputs while circumventing the limitations often associated with GAN-based methodologies. Notably, diffusion-based techniques have proven effective in various image generation tasks, including super-resolution [6, 19, 28], inpainting [32, 33], and image translation [23, 29]. However, the unique challenge of (BFR) introduces an additional layer of complexity - recovering the original identity, which may not always align with conventional measures of perceptual quality. Balancing the task of hallucinating the finer details while preserving or restoring subtle, identity-specific facial features proves challenging within the standard DDPM framework, as it lacks any identity-preserving regularization by default. To tackle this challenge, we introduce a secondary network, the Identity Preserving Conditioner (IPC), dedicated to recovering identity-specific finer facial details. Once this valuable ID information is retrieved, we employ it to condition the reverse diffusion process, ensuring the preservation of these critical details throughout the generation process.

We formulate the blind-face-restoration task as a conditional generative process, where we iteratively produce a restored face image in the pixel space, conditioned on both the degraded input and the output of IPC. Following [28]; the denoising UNet takes the degraded input as a conditional input during training. Next, to constrain the stochastic reverse diffusion process and prevent unwanted identity alteration while trying to generate sharper features, we introduce a gradient-based guidance using the output of IPC. At each step, we update the reverse trajectory towards the direction of the recovered identity by updating the diffusion score, utilizing a pre-trained face recognition model.

Unlike conventional classifier guidance [10], which uniformly adjusts scores across all pixel locations, we've observed that for BFR, such uniform regularization can undermine perceptual quality. Although IPC excels in recovering identity-specific facial regions like eyes and mouth, it may lack finer facial details in other areas, which could inadvertently impact the diffusion network's output. To address this, we introduce a learnable spatial mask that identifies crucial facial components and selectively updates scores only in those areas, minimizing unnecessary modifications elsewhere. We demonstrate that our strategy can generate visually pleasing output while improving the identity score significantly. We also use the output of IPC to initialize the reverse diffusion process instead of starting from pure noise. In contrast to [7,40], which primarily emphasize expediting the reverse diffusion process, we demonstrate through our experiments that incorporating identity information into the initial estimate can substantially enhance reconstruction fidelity in the context of blind face restoration.
To summarize, our main contributions are

- We deploy two specialized networks to disentangle the two objectives of BFR: generating visually pleasing results and recovering the underlying identity. Specifically, we use an IPC network that primarily recovers identity-specific fine-grained features from a degraded face image. Next, preserving this information, we formulate a conditional diffusion process to generate the final output with high perceptual quality.
- We propose a region-adaptive regularization strategy for the reverse diffusion process, where we selectively update the score function of just the facial areas essential for the identity information. We utilize a learnable binary mask that automatically identifies the crucial pixel locations to steer the reverse diffusion toward the recovered identity while allowing the unconstrained generation of sharper details for the remaining regions. This strategy enables a better balance of perceptual quality and face-recognition performance.
- We demonstrate the superiority of our hybrid approach through extensive experiments on multiple real and synthetic datasets.

2. Related Works

Various methods have been proposed to handle the ill-posedness of the BFR task. Facial landmarks [5, 17, 43], parsing maps [4, 30], facial heatmaps [3] were used to improve the performance. Reference-based approaches [11, 20, 21] usually need reference images with the same identity as the degraded input, which is difficult to satisfy. These priors require estimations from the corrupted images,difficult for complex real-world cases. Recent works usually utilize generative priors using a pre-trained high-quality face generation model. These methods optimize the latent vector for GAN inversion techniques [12,24] or direct projection of the input image to the latent space [27]. [39], and [34] exploited the generative prior inside an encoder-decoder framework, with structural details from the degraded input through skip connections. But, highly compressed latent space often results in loss of finer details.
Very recently, diffusion and score-based models have shown improvement over generative-prior-based works. An iterative refinement strategy has been adopted by [28], [36] for super-resolution and motion deblurring tasks. [6] used a pre-trained diffusion model and guided the reverse process with low-frequency information from a conditional image. However, such a conditioning strategy does not translate well for the BFR problem with high degradation and may alter the identity. [40] uses an unconditional diffusion model and starts from an intermediate stage of the reverse diffusion process using the output of a deterministic network. But, as the underlying diffusion model is unconditional, the restored face changes considerably compared to the original person if the reverse process is run longer. If it is used for a smaller timespan to reduce the identity alteration, the visual quality and sharpness of the output suffer considerably. [7] addressed only non-blind image super-resolution tasks, and its identity-preserving capability is yet to be tested for more difficult blind face restoration scenarios. Our work mainly focuses on balancing identity preservation and facial details restoration using a region-adaptive diffusion approach.

2.1. DDPM

To train a BFR model, we start with a paired dataset $D = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$. $\boldsymbol{x}_i$ denotes the degraded face corrupted by a complex combination of various factors, such as blur, noise, low-resolution, encoding artifacts, and $\boldsymbol{y}_i$ represents the corresponding clean face image and, N is the total number of samples. Our goal is to learn a parametric approximation to $p(\boldsymbol{y}|\boldsymbol{x})$ which is the conditional distribution of a clean image ($\boldsymbol{y}$) given a degraded image ($\boldsymbol{x}$), using a conditional DDPM model; similar to [28]. In the forward diffusion process $q(.)$, we gradually add Gaussian noise to a clean image $\boldsymbol{y}_0$ for T time steps leading to a T-step Markov chain $\boldsymbol{y}_0, \boldsymbol{y}_1, ..., \boldsymbol{y}_T$. In the reverse process $p(.)$, we denoise starting from a pure Gaussian noise $\boldsymbol{y}_T \sim \mathcal{N}(0, \boldsymbol{I})$ and iteratively refine it to obtain a clean estimate. [14] has simplified this formulation to closed-form expressions. The relevant equations are

$$q(\boldsymbol{y}_t|\boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{y}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{1}$$

$$p(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{y}_{t-1}; \boldsymbol{\mu}(\boldsymbol{y}_t, \boldsymbol{y}_0, \alpha_t), \sigma^2\boldsymbol{I}) \tag{2}$$

where $\bar{\alpha}_t = \prod_{j=1}^{T} \bar{\alpha}_j$, α_t is known as the noise schedule controlling for the diffusion process.

However, the reverse diffusion (Eq. 2) itself requires $\boldsymbol{y}_0$ that we are trying to generate. For a conditional setup, to address this, [28] utilized a denoising network to approximate $\boldsymbol{y}_0$ from $\boldsymbol{y}_t$ and $\boldsymbol{x}$. We can train a network f_θ that takes $\boldsymbol{y}_t$ and $\boldsymbol{x}$ as input and produces $\hat{\boldsymbol{y}}_0$. Equation 1 can be reformulated in terms of noise ϵ that relates $\boldsymbol{y}_t$ and $\boldsymbol{y}_0$ as

$$\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{3}$$

3. Method

Although the reverse process of standard DDPM is conditioned on the coarse estimates of y_0 and the degraded image x, it acts as a weak identity-conditioner in the denoising process. It lacks any explicit regularizing factor for the recovered identity, which is required for the BFR task. Hence, in our approach, we integrate an identity-preserving scheme for the reverse diffusion trajectory and detail our algorithm in the subsequent sections.

3.1. Identity Preserving Conditioner (IPC)

We first generate a better estimate of the underlying identity information, using a deterministic neural network which we call identity preserving conditioner(IPC) g_ϕ before starting the reverse diffusion process. We train g_ϕ to produce an estimate of $\boldsymbol{y}_0$ from $\boldsymbol{x}$ using identity-preserving loss with a well-trained ArcFace model [9], with a small amount of standard L_1 loss for better stability as

$$\mathcal{L}_{IPC} = L_1(g_\phi(\boldsymbol{x}), \boldsymbol{y}_0) + D_{cos}(f_{arc}(g_\phi(\boldsymbol{x})), f_{arc}(\boldsymbol{y}_0)) \tag{4}$$

where D_{cos} denotes the cosine distance between two feature vectors and f_{arc} denotes a pre-trained ArcFace model. Training a network with regression loss typically produces overly smooth results without sharper and realistic facial features. But, the goal of the IPC is not to generate visually pleasing restored face images but to produce a stable approximation $\hat{\boldsymbol{y}}_0$ while recovering the identity information as much as possible. Moreover, from the qualitative results in section 4.4 observe that although $g_\phi(\boldsymbol{x})$ fails to recover intricate high-frequency details of the face like skin texture or hair patterns, it mainly focuses on recovering the critical facial areas such as eyes, nose, mouth, etc., which are most important for the identity.

Note that our design differs from [40], which simply employs a coarse estimation network for faster sampling in the diffusion process. Such a network, trained without explicit identity preserving objective, fails to recover the underlying identity information satisfactorily, as validated in our experimental section. Similarly, our work also differs from [35], which utilizes both adversarial and identity-preserving loss in a single network and uniformly for all the pixel locations, which is difficult to optimize for the challenging BFR task. Our disentangled and region-adaptive approach preserves a better balance between the perceptual quality and the recovered identity.

3.2. Identity Preserving Conditional Diffusion

We optimize the parameters of our Diffusion model using

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{\boldsymbol{y}_0, \boldsymbol{x}, \epsilon, \bar{\alpha}} ||\epsilon - f_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + (1-\bar{\alpha}_t)\epsilon, g_\phi(\boldsymbol{x}), \bar{\alpha})|| \tag{5}$$

During inference, we use the forward process (Eq. 3) to generate a noisier version of $g_\phi(\boldsymbol{x})$, i.e., $\hat{\boldsymbol{y}}_{T'}$, in a single step. Basically, $p_\phi(\hat{\boldsymbol{y}}_{T'}|\boldsymbol{x})$ is an approximation of $p(\boldsymbol{y}_{T'}|\boldsymbol{y}_0)$. Next, we start the reverse diffusion process from $\hat{\boldsymbol{y}}_{T'}$. [40] demonstrate that the estimation error in $p_\phi(\hat{\boldsymbol{y}}_{T'}|\boldsymbol{x})$ reduces as we increase T'. Although, the absolute error is reduced, we observe that typically there is a trade-off between the perceptual quality of the recovered face and the preservation of facial features that were present in $\boldsymbol{x}$ and $g_\phi(\boldsymbol{x})$. As [40] utilizes an unconditional diffusion model, running it for a smaller time step usually results in lower visual quality but higher identity preservation. If we run it longer to improve the restoration quality, the identity of the face is significantly altered. This effect is more prominent when the level of degradation is high in the input image. Thus, we focus on building a framework to achieve better perceptual quality without compromising identity-specific facial features. Unlike [40], that utilizes an

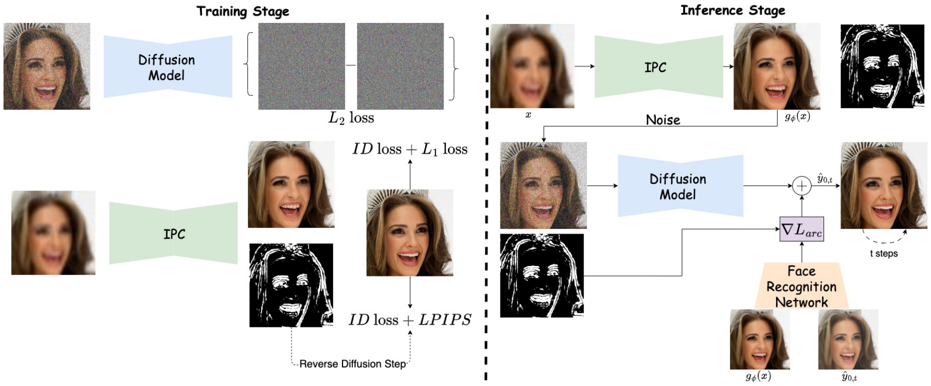

Figure 1. An overview of our approach. On the left, we show the training stage, where the diffusion model is trained using standard denoising loss, and the IPC network is trained with regression and identity loss. During inference (right), we update the diffusion score using the gradient calculated from the recognition loss, estimate of IPC, and the intermediate output of reverse diffusion.

unconditional model, we use a conditional diffusion model where the denoising UNet takes the degraded image $\boldsymbol{x}$ as an additional input. Although, it allows a longer reverse diffusion process without adversely affecting the output, such weak supervisions are still sub-optimal in preserving the identity information. To this end, we use an identity preserving gradient-based guidance strategy at each time step constraining the trajectory of the iterative refinement process.

[10] used a classifier pretrained on noisy images to guide the generation process toward a target class. Similar strategies have been observed for the text-guided image generation tasks [18] as well. In our case, we utilize a standard face recognition model ArcFace [9] pretrained on clean images. Instead of training it to recognize noisy face-images, which is a difficult task by itself, we denoise the noise image at any timestep by utilizing the inherent capability of the diffusion UNet to denoise different noise levels. Recall that that the denoising neural network f_θ is trained to produce the noise that was added to $\boldsymbol{y}_0$ to produce $\boldsymbol{y}_t$ (Eq. 3). Thus, at every reverse time step $t < T'$, we can approximate $\hat{\boldsymbol{y}}_{0,t}$ by rewriting Eq. 3 as

$$\hat{\boldsymbol{y}}_{0,t} = \frac{\boldsymbol{y}_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{(1-\bar{\alpha}_t)}}{\sqrt{\bar{\alpha}}}\epsilon \tag{6}$$

As our IPC is trained using identity preserving loss, it mainly focuses on recovering identity-specific details as much as possible from the degraded image. The estimated $g_\phi(x)$ works much better for the face-recognition model than the image generated by a standard diffusion model, which is trained with a different objective function altogether. Thus, we utilize the identity features recovered in $g_\phi(\boldsymbol{x})$ to guide the reverse diffusion process. An ArcFace-based loss $\mathcal{L}_{arc}$ can be defined as

$$\mathcal{L}_{arc} = D_{cos}(f_{arc}(\hat{\boldsymbol{y}}_{0,t}), f_{arc}(g_\phi(\boldsymbol{x}))) \tag{7}$$

At each time step, we calculate the gradient w.r.t $\boldsymbol{y}_t$ and update the default score function in the reverse diffusion process (Eq. 2) towards the direction of minimizing $\mathcal{L}_{arc}$.

3.2.1 Region adaptive Masking Scheme

Although the method discussed above improves recognition accuracy, we observe that it adversely affects the overall perceptual quality of the restored image. We suspect that the underlying objective of a recognition model and the need to generate visually pleasing facial details in the BFR task need not align for all spatial locations. For example, a recognition model might work well when the identifying facial features, such as the eye, nose, etc., are adequately reconstructed even if the other regions are not sharp enough, as it is the case for IPC. Thus, to keep a balance between the two, we use a region-adaptive gradient-guidance strategy, where we update the score function in the reverse diffusion step only for those spatial regions in the face which contribute most to the recognition performance. Instead of selecting such regions in a handcrafted manner, we add a small sub-branch to IPC that predicts a spatial binary mask $\boldsymbol{M}$ depicting the regions crucial for recognition. As the IPC model is explicitly trained to focus on recovering identity-specific features, we extract a proxy information $\boldsymbol{M}'$ from its output by taking the difference between its output and the degraded input followed by a thresholding operation as follows.

$$\boldsymbol{M}' = \begin{cases} 1 & |g_\phi(x) - x| > \delta \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

These regions ($\boldsymbol{M}'$) with significant changes, along with the original degraded image $\boldsymbol{x}$ is fed to the mask prediction sub-branch as input, to ease the learning process. We have visualized some of the facial masks in Fig. 6. We use a threshold of 0.3 in our work. Next, at a random time step t, we use the mask $\boldsymbol{M}$ to apply the gradient on certain locations and produce $\boldsymbol{y}_{t-1}$. At each step, we update the parameterized mean $\boldsymbol{\mu}(\boldsymbol{y}_t, \boldsymbol{y}_0, \alpha_t)$ of $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})$ as

$$\boldsymbol{\mu}'(\boldsymbol{y}_t, \boldsymbol{y}_0, \alpha_t) = \boldsymbol{\mu}(\boldsymbol{y}_t, \boldsymbol{y}_0, \alpha_t) + (\nabla_{\boldsymbol{y}_t} \mathcal{L}_{arc}) \odot \boldsymbol{M} \quad (9)$$

The mask prediction sub-branch is trained to optmize the face recognition loss (Eq. 7) and perceptual loss (LPIPS) w.r.t $\hat{\boldsymbol{y}}_{0,t}$ at different time steps. Our region-adaptive strategy improves the identity-preserving property of the restored image without undermining the visual quality. An alternative could be to try a separate face-parsing network, but this would increase the computational load.

Table 1. Quantitative evaluation on 3000 images of size 256×256 from the CelebA-Test (BFR). Bold and underline indicate the best and the second best performance.

Methods	LPIPS ↓	FID ↓	IDS ↑	PSNR ↑	SSIM ↑
GPEN	0.3336	120.46	0.3964	22.49	0.6074
GFPGAN	0.2783	110.52	0.4607	22.56	0.6125
PSFRGAN	0.2513	88.75	0.3934	21.75	0.5450
CodeFormer	0.2322	56.00	0.4955	22.39	0.5778
DifFace	0.2028	70.69	0.4808	22.82	0.6190
RestoreFormer	0.2907	60.98	0.3982	21.77	0.5301
IPC	0.3109	118.50	**0.5849**	24.16	0.6826
IPC w/o ID Loss	0.3344	127.53	0.5155	**24.26**	**0.6921**
Ours	**0.1898**	**55.42**	0.5415	22.34	0.6087

Table 2. Quantitative evaluation on 3000 images of size 512×512 from the CelebA-Test (BFR). Bold and underline indicate the best and the second best performance.

Methods	LPIPS ↓	FID ↓	IDS ↑	PSNR ↑	SSIM ↑
GPEN	0.3362	101.12	0.4022	22.43	0.6009
GFPGAN	0.2812	99.03	0.4633	22.50	0.6060
PSFRGAN	0.2513	64.81	0.3983	21.75	0.5450
CodeFormer	0.2288	54.41	0.5009	22.35	0.5736
DifFace	0.2061	52.18	0.4833	**22.74**	**0.6116**
Ours	**0.1966**	**48.12**	**0.5654**	21.42	0.5612

4. Experimental Results

4.1. Training Dataset

The FFHQ dataset [16] contains 70000 high-quality face images of resolution 1024×1024. For training, we resized all the images to 512×512, considered as the GT. Since our approach requires degraded-clean pairs training, we synthesize degraded images on the FFHQ dataset using the degrading model proposed in [4, 34, 35, 39] as

$$\boldsymbol{I}_{deg} = ((\boldsymbol{I} \otimes k) \downarrow_s + \boldsymbol{n}_\sigma)_q \quad (10)$$

where $I, I_{deg}, k, n_\sigma, s, q$ are the clean face image, corresponding degraded image, the blur kernel, the Gaussian noise with a standard deviation σ, downscaling factor and the JPEG-compression quality factor, respectively. In our implementation, we sample σ, s, q randomly and uniformly from [0,20], [1,32], and [30,90]. Finally, the degraded image is resized back to 512×512.

4.2. Testing Dataset

We first evaluate our approach on a synthetic dataset CelebA-Test for the BFR task, which contains 3000 images selected from the CelebA-HQ dataset [15], where the degraded images are synthesized under the same degradation range as our training settings. Further, we test our method on real-world datasets: WebPhoto-Test [34], WIDER Face [38] (970 images), and TURB. WebPhoto-Test consists of 407 low-quality faces extracted from the internet. We also evaluate on images affected by atmospheric turbulence from the BRIAR [8] and LRFID dataset [25]. We randomly sample 139 images with different identities from these datasets for the TURB dataset, which provides a more challenging scenario, as the models were never trained on severe turbulence-affected images.

Table 3. Quantitative evaluation on 3000 images from the CelebA-Test for extreme upsampling from 16×16 images ($\times 32$). Bold and underline indicate the best and the second best performance, respectively.

Methods	LPIPS ↓	FID ↓	IDS ↑	PSNR ↑	SSIM ↑
GPEN	0.4350	148.39	0.1843	19.91	0.5346
GFPGAN	0.4028	160.29	0.2243	19.95	**0.5366**
CodeFormer	0.3565	73.45	0.2546	19.14	0.4639
DifFace	0.3001	53.93	0.2892	**20.12**	0.5314
RestoreFormer	0.4193	103.13	0.1438	19.26	0.4581
Ours	**0.2885**	**48.76**	**0.3376**	19.08	0.5194

Table 4. Quantitative comparisons of FID (↓) on real-world datasets in terms of FID.

Methods	WIDER Face	WebPhoto	CelebA-Child	TURB
PSFRGAN	49.85	88.45	107.40	147.54
GPEN	46.99	81.77	109.55	166.22
GFPGAN	39.76	87.95	111.78	161.14
CodeFormer	39.21	116.18	116.18	126.55
DifFace	37.49	85.52	110.81	133.86
Ours	**35.56**	**81.19**	**104.40**	**123.01**

4.3. Evaluation Metrics

For quantitative evaluation, we mainly focus on Frechet Inception Distances (FID) [13] and Learned Perceptual Image Patch Similarity (LPIPS) [41] metric, as these correlate better with the perceptual quality and realness of the restored images. For completeness, we also calculate the PSNR and SSIM values, which often fail to adequately reflect the visual quality. To measure the face recognition performance, we follow [34, 35] and calculate the cosine similarity between the features of the restored image and

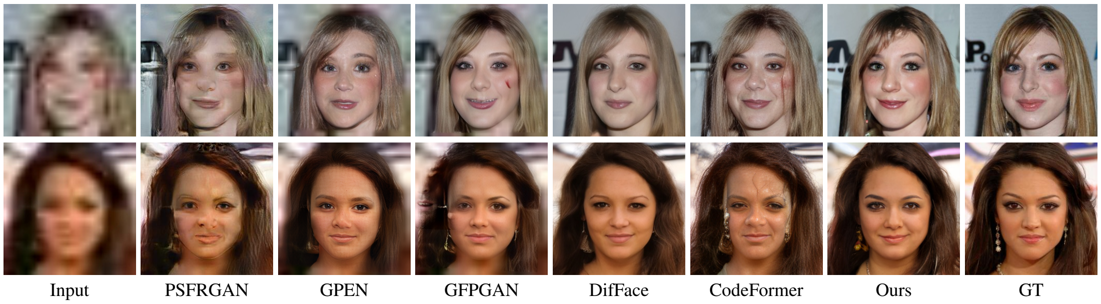

Figure 2. Qualitative comparisons on CelebA-Test set for BFR.

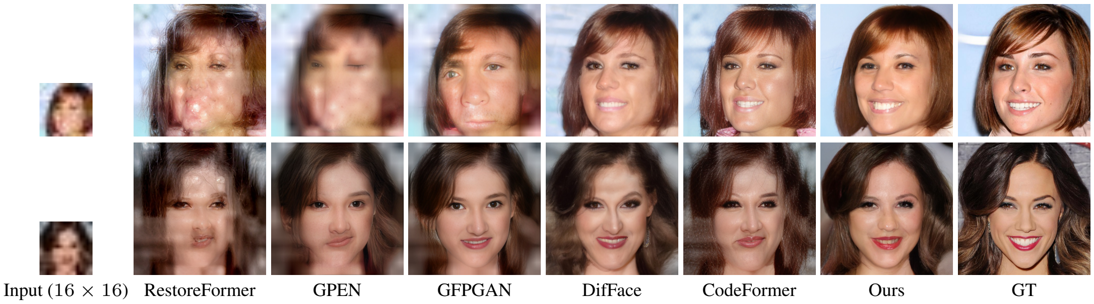

Figure 3. Qualitative comparisons on CelebA-Test set for ×32 upsampling. Although the input is severely degraded and contains minimal information, our approach works better than existing approaches in restoring the face. GT represents the ground truth

the corresponding paired GT image (IDS). Higher cosine similarity indicates better identity preservation and recovery in the output image. We use the same evaluation protocol and pretrained-models for LPIPS, FID and ArcFace as prior art [34, 35]

4.4. Comparisons with State-of-the-Art Methods

We compare quantitatively and qualitatively with the following state-of-the-art (SOTA) methods: Pulse [24], PSFRGAN [4], GPEN [39], GFPGAN [34], CodeFormer [42], RestoreFormer [35] and DifFace [40]. We use the official model and results provided by the authors for comparison.

Synthetic BFR: First, we measure the restoration accuracy on the synthetic CelebA-Test dataset. To verify the robustness of the restoration algorithms, we perform this experiment under various settings. Existing works are usually trained on 512 × 512 images and also require inputs of the same dimension for testing. In real life, face images are often of lower resolution. Thus, we created a synthetic dataset of size 256 × 256 and used that for evaluation. All the inputs were resized to 512 × 512 using simple interpolation before feeding to existing works, but such inputs are of lower quality than images of 512 dimension. The quantitative results w.r.t. ground-truth images of the same size are reported in Table 1. As we can observe, our IPC model achieves the highest IDS score as it is directly trained to optimize that objective. In comparison, existing GAN, VAE, or diffusion models achieve suboptimal scores as they struggle to recover the identity while generating visually pleasing information. In contrast, our approach can achieve a much better balance between the IDS and other perceptual metrics, such as LPIPS and FID.

Next, we perform the same experiment on 512 × 512 images from the CelebA-Test. The results are reported in Table 2. We observe a similar trend in this setting too. Most notably, our approach achieves a significant boost over the prior art for identity recovery, demonstrating our approach's superior ability for faithful restoration of face images, even at higher resolution.

Extreme Upscaling: Next, we test the algorithms under an extreme setting of BFR. We set a fixed downscaling factor of ×32 to 512 × 512 images to generate degraded images of size 16 × 16. To make it more challenging, we further introduce a certain amount of noise and blur to the images. The quantitative results are reported in Table 3. Although the overall IDS of all the algorithms are considerably lower as the input images have very limited information under such difficult settings, our approach still achieves comparatively better perceptual quality as well as recognition accuracy.

Real-World BFR: We further compare our approach on a range of real-world degraded datasets. As the GT pairs are not available, we report the FID score in Table 4 and visualized the images in Fig. 4. Existing GAN-based approaches like GPEN, GFPGAN, etc., often produce overly

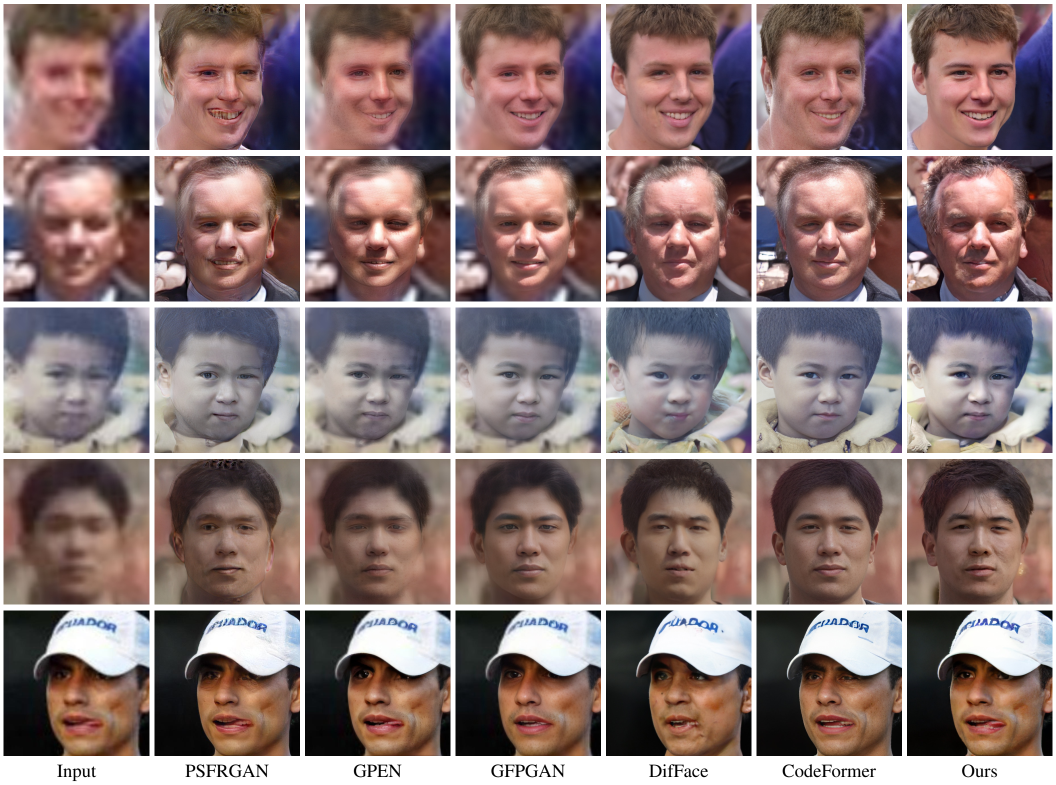

Figure 4. Qualitative comparisons on real-world dataset. The first two rows represent images from WIDER Face dataset, the next two rows from the WebPhoto dataset and the last row contains image from the LFW datasets, respectively.

smooth output. CodeFormer output has visible artifacts (Row 1, Fig. 4) or repetitive skin/hair texture (Rows 2,3, Fig. 4). DifFace outputs are less sharp and even slightly alter the facial details as they use an unconditional diffusion model. Our approach produces a more realistic and faithful reconstruction with fewer artifacts, even for low to medium degradation.

Face Recognition: We also measure the face-recognition accuracy on the TURB dataset using the pretrained ArcFace network. We report the top-1/3/5 recognition accuracy (w.r.t gallery images with the same identity) in Table 5. As can be observed, our restoration approach significantly boosts the performance of downstream face-recognition tasks on real-world dataset as well.

5. Ablation Analysis

Table 5. Face recognition accuracy using pre-train ArcFace [9] on real-world BRIAR/*LRFID* dataset. Our method performs best for such downstream task as well.

Methods	Top-1 (↑)	Top-3 (↑)	Top-5 (↑)
GPEN	32/*50.6*	50/*72*	62/*81*
GFPGAN	26/*57*	58/*79*	60/*85*
CodeFormer	28/*61*	52/*77*	58/*82*
RestoreFormer	20/*62*	50/*81*	62/*88*
Ours	**38**/***68***	**62**/***88***	**70**/***94***

Table 6. Quantitative comparison of different ablations of our network on a subset of CelebA-Test with images of dimension 256×256. IPC, ID and M-ID represents: using identity preserving conditioner network, gradient-based guidance and region-adaptive gradient-based guidance, respectively.

Methods	IPC	ID	M-ID	LPIPS ↓	FID ↓	IDS ↑
Net1				0.2681	**70.56**	0.48
Net2	✓			**0.1854**	71.2	0.51
Net3	✓	✓		20.39	82.10	**0.60**
Net4	✓		✓	0.1862	72.17	0.58

In Table 6, we analyze the effect of individual components of our approach on the perceptual quality and identity-preserving aspect. We use a subset of CelebA-Test with 256×256 images for our ablation. We use the same U-Net architecture from Guided-Diffusion (GD) [1] as the denoising model except for additional downsampling and upsampling layers to handle 512×512 without significantly increasing the computational cost. We empirically observed that the IPC is model agnostic, and any SOTA restoration network

[1] https://github.com/openai/guided-diffusion

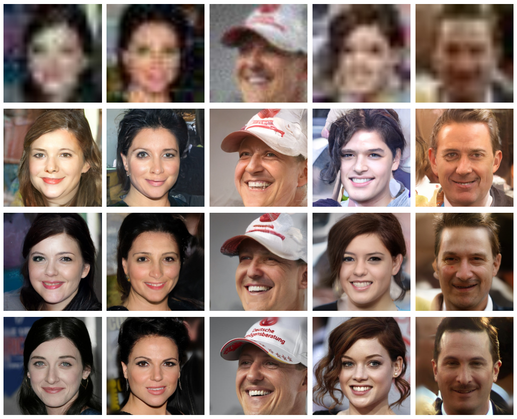

Figure 5. Comparisons between Guided-Diffusion (GD) strating from noise with 1000 steps, starting from the output of IPC with 400 steps (without region-adaptive gradient guidance). From top: input, o/p of GD (t=1000), o/p of GD + IPC (t=400), ground-truth.

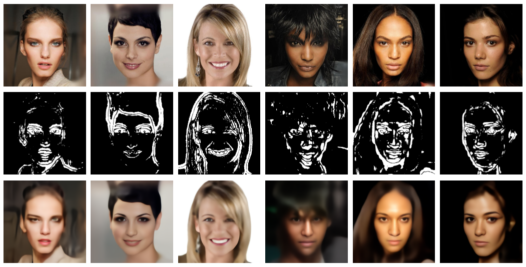

Figure 6. Visualization of mask M (Row 2), highlighting the crucial facial locations and the output of IPC (Row 3).

results in a comparable identity-recovering accuracy. We intuit that given a strong backbone, the objective function (Eq. 4) plays a significant role in the performance of IPC. We finally select SwinIR [22] that keeps a good balance between performance and accuracy.

Net1 is our backbone diffusion model without using IPC or gradient-based guidance at inference. The framework is similar to GD except for the additional downsampling layer. For Net1, we perform the reverse diffusion process for 1000 time steps starting from pure Gaussian-noise, similar to GD. We observe that when the input degradation is severe, starting from pure noise while using the corrupted image as a condition often leads to unwanted artifacts or changes in the restored image, as shown in Fig. 5. The same is also reflected in the quantitative performance. Next, we include the IPC in the reverse diffusion process to have an initial estimate and run the reverse process for 400 timesteps. It achieves a much more stable performance across a wide range of degradations, supported by the improved quantitative performance in Table 6. Although it improves the perceptual quality of the restored image, we observed that it is suboptimal in preserving the identity-specific features as the IDS score drops compared to the output of the IPC. To address this, in Net3, we include gradient-based guidance using a well-trained face recognition network. Specifically, we use the identity-specific features recovered in the output of the IPC to shift the intermediate output of the reverse diffusion process in the direction of recognition-accuracy improvement. Although it significantly boosts the IDS score, it adversely affects perceptual quality. We suspect that, as the gradient signal w.r.t to the identity of the IPC estimate is not 100% ideal, small unwanted perturbations in the spatial regions, which are not crucial for the recognition performance, may harm the restoration quality. Thus, we utilize region-adaptive guidance in Net4, our final model. The binary mask $\boldsymbol{M}$ identifies the crucial facial regions or pixels essential for the face-recognition performance. Thus, in Net4, we use gradient-based steering only for those regions, avoiding unwanted perturbations in the rest of the areas and preserving the overall perceptual quality and sharpness.

Time(s) required and FID scores (under the same settings) of GFPGAN, CodeFormer, DifFace, Guided-Diffusion and Ours are : 0.5/107,0.15/64,6/63,30/58,12/59. Diffusion models have much better FID but slower runtime. It is an active research area, and in future, more efficient models, such as [31], could potentially be used, but is beyond the scope of the current work. We also visualize the outputs and the binary masks predicted by IPC in Fig. 6.

6. Conclusions

We propose a region-adaptive diffusion model to restore severely-degraded face images. Our design achieves a better balance between restoration quality, identity recovery, and efficiency than existing diffusion-based approaches. Despite the significant improvement, the identity-preserving ability of our approach is limited by the performance of the IPC network. In future, a better IPC model, recognition model, or techniques to decide whether the recovered identity information is correct can improve the performance further. Our reverse diffusion process can also be further accelerated by adaptively selecting the intermediate time step to start from, which we will explore in the future.

7. Acknowledgement

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U. S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Simon Baker and Takeo Kanade. Hallucinating faces. In *Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, pages 83–88. IEEE, 2000. 1

[2] Thirimachos Bourlai, Arun Ross, and Anil K. Jain. Restoring degraded face images: A case study in matching faxed, printed, and scanned photos. *IEEE Transactions on Information Forensics and Security*, 6(2):371–384, 2011. 1

[3] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2018. 2

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11896–11905, 2021. 2, 5, 6

[5] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018. 2

[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 1, 2

[7] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022. 2

[8] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023. 5

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3, 4, 7

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2, 4

[11] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[12] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020. 2

[13] 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. *Advances in neural information processing systems*, 30, 2017. 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[17] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019. 2

[18] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 4

[19] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1

[20] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 399–415. Springer, 2020. 2

[21] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018. 2

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 8

[23] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1

[24] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 1, 2, 6

[25] Kevin J Miller, Bradley Preece, Todd W Du Bosq, and Kevin R Leonard. A data-constrained algorithm for the emulation of long-range turbulence-degraded video. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXX*, volume 11001, pages 204–214. SPIE, 2019. 5

[26] Masashi Nishiyama, Hidenori Takeshima, Jamie Shotton, Tatsuo Kozakaya, and Osamu Yamaguchi. Facial deblur inference to improve recognition of blurred faces. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1122. IEEE, 2009. 1

[27] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 1, 2

[28] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 3

[29] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 1

[30] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8260–8269, 2018. 2

[31] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 8

[32] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[34] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 1, 2, 5, 6

[35] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 1, 3, 5, 6

[36] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 2

[37] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1551–1560, 2020. 1

[38] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 5

[39] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 2, 5, 6

[40] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 2, 3, 6

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[42] Shangchen Zhou, Kelvin CK Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *arXiv preprint arXiv:2206.11253*, 2022. 6

[43] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 614–630. Springer, 2016. 2