Rethinking Deep Face Restoration

Yang Zhao*
University at Buffalo
yzhao63@buffalo.edu

Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu
Google Research
{ycsu, ctchu, yandongli, renn, yukun}@google.com

Changyou Chen
University at Buffalo
changyou@buffalo.edu

Xuhui Jia
Google Research
xhjia@google.com

Abstract

A model that can authentically restore a low-quality face image to a high-quality one can benefit many applications. While existing approaches for face restoration make significant progress in generating high-quality faces, they often fail to preserve facial features that compromise the authenticity of reconstructed faces. Because the human visual system is very sensitive to faces, even minor changes may significantly degrade the perceptual quality. In this work, we argue that the problems of existing models can be traced down to the two sub-tasks of the face restoration problem, i.e. face ***generation*** *and face* ***reconstruction****, and the fragile balance between them. Based on the observation, we propose a new face restoration model that improves both* ***generation*** *and* ***reconstruction****. Besides the model improvement, we also introduce a new evaluation metric for measuring models' ability to preserve the identity in the restored faces. Extensive experiments demonstrate that our model achieves state-of-the-art performance on multiple face restoration benchmarks, and the proposed metric has a higher correlation with user preference. The user study shows that our model produces higher quality faces while better preserving the identity* 86.4% *of the time compared with state-of-the-art methods.*

1. Introduction

Face images play a critical role in our daily life and are at the very center of success for many applications such as portrait taking, face identification, etc. While these applications usually rely on having decent quality faces as inputs, low-quality face images are inevitable in the real world due to various reasons, e.g. low image resolution, motion blur, defocus blur, sensor noises, encoding artifacts, etc. Therefore, a method that can faithfully restore a degraded face into a high-fidelity one regardless of the type of degradation is highly desired.

Much progress has been made in face restoration in the past few years, thanks to the rapid development of deep generative adversarial networks (GANs) [8]. Existing works treat face restoration as a conditional image generation problem, and they learn a U-Net model that predicts a high-quality face image given a low-qualiy one as input [3, 20, 21, 23, 35, 36, 40]. Despite being able to generate realistic faces, they still suffer from unique challenges introduced by face restoration. Specifically, they often fail to preserve delicate facial features in the input but instead hallucinate a high-quality face that does not resemble the original subject. The model may change the subject's eye color, skin texture, shape of face components, etc, as shown in Figure 1. While these changes may be negligible in pixel space and are irrelevant to the realisticness, they are essential for authenticity and can significantly impact downstream applications. For example, they may break a face identification system because the biometric characteristics deviate from the original subject, and they may degrade the perceptual quality of a photo because the subject looks like a different person.

We argue that the above issues are caused by the fragile balance between face generation and face reconstruction. As we will show later, the face restoration problem can be interpreted as a combination of two sub-tasks, i.e. ***generation*** and ***reconstruction***, where face ***generation*** aims to learn the distribution of high quality faces and face ***reconstruction*** aims to capture the face characteristic (e.g. shape and texture) from an image regardless of its quality [5, 36]. A model that overemphasizes generation and fails in reconstruction may hallucinate a face that does not belong to the subject. In contrast, a model that fails in generation leads to unsatisfactory restoration quality. Therefore, a successful face restoration model has to address the two sub-tasks simultaneously, which remains to be realized.

Based on the observation, we propose a new model that aims to improve both generation and reconstruction. To

*Work done during an internship at Google Research.

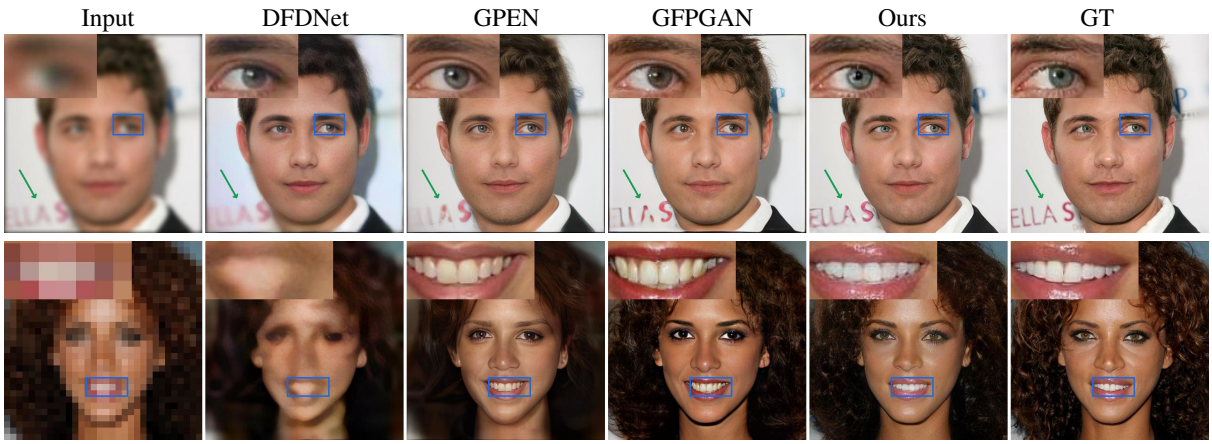


Figure 1. Problems of state-of-the-art face restoration models. GPEN [40] and GFPGAN [36] are biased toward face generation and may alter facial details (e.g. eye color) that are highly correlated with the identity. DFDNet [20] is biased toward reconstruction and does not remove all degradations. Our approach achieves the best balance and restore a high quality face while preserving the identity.

improve face generation, we inject an adaptive conditional noise to the model, motivated by the great success of recent image generation models. The noises empower the restoration model with stochastic property and allow the model to capture the non-deterministic nature of the face restoration problem. To improve face reconstruction, we enhance the latent features in the skip connections by 1) quantizing the features using a codebook learned from high-quality images and 2) introducing a global feature fusion module for an adaptive combination of the features from the decoder and the skip connections. These improvements are based on the observations that the features extracted by the encoder may harm the reconstruction performance, especially when the input quality is poor. Finally, we explore the model architecture, particularly the number of skip connections, to optimize the balance between generation and reconstruction.

Like the models, the evaluation metrics for face restoration also suffer from overemphasizing either the generation or the reconstruction aspect of the problem. Existing works borrow either metrics designed for image generation, e.g. Fréchet inception distance (FID) [12], or metrics developed for image reconstruction, e.g. Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), or Learned Perceptual Image Patch Similarity (LPIPS) [44]. They focus on the perceptual quality or the pixel similarity between the output and the target respectively, and neither of them was able to capture subtle changes in facial features. To this end, we propose a new metric that measures both image quality and content preservation, where content preservation is defined by the ability to preserve the identity. Experiment results demonstrate that the proposed metric better correlates with the perceptual quality of human raters in the face restoration problem.

The main contributions of this paper are as follows. First, we show that issues of existing face restoration models may be traced down to the two sub-tasks of the problem, i.e. face generation and face reconstruction. Second, we propose a new face restoration model by improving the model design for both sub-tasks. Finally, we introduce a new evaluation metric for face restoration that measures both the perceptual quality and identity preservation. Empirical results on two benchmarks, blind face restoration (BFR) and super-resolution (SR), show that proposed model consistently outperforms state-of-the-art methods, and the proposed metric better correlates with the perceptual quality of human raters. In addition, user study shows that our model is preferred by human raters 86.4% of the time compared with state-of-the-art face restoration models.

2. Related work

Face restoration Face image restoration has attracted considerable attention from various aspects, *e.g.,* face super resolution [11, 22, 23, 37, 39], blind face restoration [20, 21, 36, 40], deblurring [18, 32, 41], denoising [10, 43], inpainting [38, 42, 47], etc. Human perceptions are more sensitive to facial images than other image domains and thus demand more concrete and meticulous control. In terms of modeling strategy, all recent notable works on high-resolution (*e.g.,* 512×512) resort to maximum likelihood estimation to reconstruct realistic face characteristics and adversarial learning to generate a high-fidelity image distribution.

State-of-the-art BFR models exploit off-the-shelf generative networks like StyleGAN [16] to improve the restoration performance [9, 23, 29, 36, 40]. Based on the assumption that the prior generative network can produce arbitrary high-fidelity faces, they focus on mapping the degraded faces into the appropriate latent features for the generator. Although they show promising performance in terms of image generation metrics, subjective evaluation shows that the models are dominated by the prior generative networks even after finetuning, leading to unfaithful restoration such as color shift or excessive hallucination. In other words, they bias toward face generation and downplay face recon-

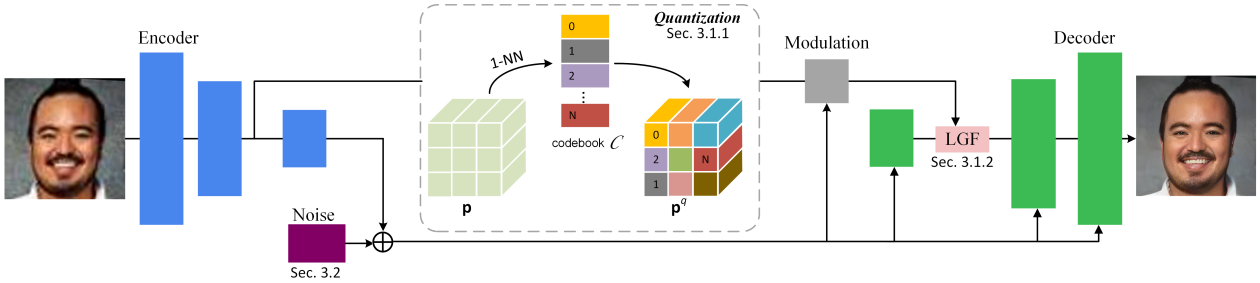


Figure 2. The proposed model with one skip connection. (1-NN: 1-nearest neighbor search. Modulation: feature modulation as in StyleGAN2 [16]. LGF: linear gated feature fusion.)

struction. In contrast, our approach reaches a good balance between content preservation and high-fidelity faces generation, which leads to a better subjective quality. Furthermore, our model can be trained from scratch and does not require a carefully optimized GAN model.

Evaluation metric Existing works for face restoration adopt PSNR, SSIM, and LPIPS [44] to measure the reconstruction performance for every example. To evaluate the distance between the restored face distribution and the real face distribution, we often adopt FID, Inception score [30] and Kernel Inception Distance [2]. However, they may cause inconsistent judgment from one to another. A well-known example is that blurring images can improve PSNR and SSIM [44] but degrade other metrics. FID is affected mainly by the number of evaluation samples and may also bring unfair comparisons without prior knowledge of the evaluation system [25]. LPIPS appears to suggest a better agreement with humans, but it fails to capture concrete face identities. We propose a robust metric to simultaneously measure overall samples' realism and individual identity preservation to address these discrepancies.

3. Approach

In this section, we introduce the proposed approach for improving face restoration. We begin by formulating the face restoration problem and describing how to break it down into the combination of face generation and face reconstruction. We next introduce how to improve the reconstruction and generation sub-tasks respectively. Finally, we describe the training process.

Problem interpretation We can interpret the face restoration problem as a combination of the face generation and face reconstruction sub-task from the objective and model perspective.

Let X denote the degraded low-quality image domain, Y indicate the high-quality image domain, and P_Y imply the distribution of high-quality images. Assume that there exists a one-to-many degradation function Deg: $Y \rightarrow X$, the goal of face restoration is to learn an inverse function G: $X \rightarrow Y$ that satisfies

$$\min_G \mathcal{D}(P_{G(X)}||P_Y) + \mathbb{E}_{y\sim Y}\mathbb{E}_{x\sim Deg(y)}\kappa(G(x), y), \quad (1)$$

where $\mathcal{D}$ is a distribution distance and $\kappa(\cdot)$ is a pair-wise distance between two images.

From the objective perspective, the first term is the objective for image generation that encourages the restored images to look realistic and be indistinguishable from authentic high-quality images. While the second term is the objective for image reconstruction, which resembles the high-quality image from which the input image is degraded and preserves facial features.

From the model perspective, the decoder in G can be considered an image generation model that aims to generate realistic images from latent features. In contrast, the encoder aims to project images to appropriate latent features for reconstruction, similar to the StyleGAN encoder [29]. Unlike StyleGAN encoder, however, the encoder in the face restoration model has to be robust to the degradation in the input image in order to restore images with arbitrary quality. A common practice is to implement G using a U-Net architecture as illustrated in Figure 2 and realize the first and second half of Eq. 1 using an adversarial loss and reconstruction losses respectively.

Based on this interpretation, we next describe how to improve the generation and reconstruction sub-tasks to achieve better face restoration.

3.1. Improving Reconstruction

The face reconstruction sub-task requires fine-grained control on face details in the generated image based on the input image to achieve authentic face restoration. This is achieved by conditioning the generation model using the latent features extracted by the encoder. More specifically, the skip connections in the U-Net architecture pass low to high-level information to the decoder for an authentic reconstruction of the input face.

Although the U-Net architecture is widely adopted in prior works, our empirical results suggest that it may be sub-optimal for face restoration, particularly for inputs with severe degradation. The encoder could not extract useful features from low-quality images, and the low-quality features hindered the restoration performance. To address this issue, we propose the following improvements for the U-Net architecture.

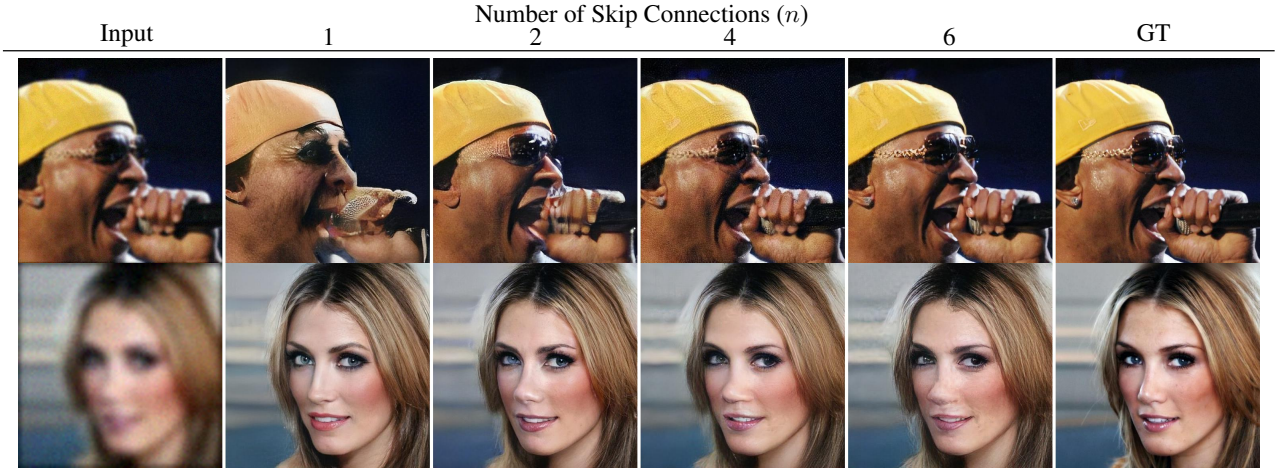


Figure 3. Qualitative comparison by varying the number of skip connections. We count from the layer with feature resolution 8×8, *i.e.,* there exist possible skip connections at resolution nodes $\{2^{n+2} \times 2^{n+2}\}_{n=1}^{6}$ when we set the maximum input resolution at 512×512.

3.1.1 Feature quantization

To help the model generalize to severely degraded images, we propose to enhance the features extracted by the encoder. In particular, we adopt the feature quantization approach that has attracted much attention in representation learning and generative model recently [6, 24, 27, 28, 46] for feature enhancement. The idea is that, given a codebook $\mathcal{C}=\{c_k\}_{k=1}^{K}, c_k \in \mathbb{R}^d$ of high quality features, we can enhance a corrupted feature $\mathbf{p}_{ij} \in \mathbb{R}^d$ by quantizing $\mathbf{p}_{ij}$ to a code word c_k in the codebook $\mathcal{C}$. In other words, we replace a feature extracted by the encoder that may be corrupted with a feature in the codebook such that the resulting quantized feature always consists of high-quality features.

We incorporate feature quantization into our model as follows. Given a learned codebook $\mathcal{C}$ and a feature map $\mathbf{p} \in \mathbb{R}^{H \times W \times d}$ extracted by the encoder, we replace the feature at each spatial location $\mathbf{p}_{ij}$ using its closest entry in $\mathcal{C}$:

$$\mathbf{p}_{ij}^{q} = \arg_{c_k} \min ||\mathbf{p}_{ij} - c_k||_2, \tag{2}$$

and the original feature map $\mathbf{p}$ is replaced by the quantized feature map $\mathbf{p}^q$ in the following operations. See Figure 2.

We learn one codebook for each skip connection feature map. During training, we optimize the following loss to encourage the model to utilize quantized features:

$$\mathcal{L}_{\text{VQ}} = ||\mathbf{p} - \text{sg}(\mathbf{p}^q)||_2^2, \tag{3}$$

where $\text{sg}(\cdot)$ is a stop-gradient operator. Instead of learning the codebook end-to-end using gradient descent, we use exponential moving average (EMA) [24, 28] over the features extracted from ground truth high-quality images to learn the codebooks. More specifically, we extract features from the ground truth high-quality images using the encoder in each iteration. We then assign each feature vector to the closest code word in the current codebook before updating the code words using the average feature. The codebooks are initialized with a normal distribution.

3.1.2 Linear gated feature fusion

Another way to address the problem of uninformative features from the encoder is to fuse only suitable features in the skip connections into the feature maps of the decoder. However, exiting works use addition, concatenation [40], or spatial feature transform [20, 36] to combine the features, and none of them are aware of whether the fused features are suitable for restoration or not. To address this issue, we propose a linear gated feature fusion (LGF) module which integrates information from both encoder and decoder to filter uninformative features. It integrates global information from both features and also filters the feature combination with a confidence score.

Let $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{H \times W \times C}$ represent the features from the corresponding encoder and decoder block respectively. The LGF module computes:

$$\begin{aligned} \text{Global score: } \mathbf{o} &= \mathsf{DownSample}_r(\mathbf{p}+\mathbf{q}) \cdot \mathbf{W} \\ \text{Gated score: } \mathbf{s} &= \mathsf{UpSample}_r(\mathsf{Sigmoid}(\mathbf{o})) \\ \text{Fused feature: } \mathbf{q}^* &= \mathbf{s} * (\mathbf{p}+\mathbf{q}) + (1-\mathbf{s}) * \mathbf{q} \end{aligned} \tag{4}$$

where r is the window size for downsample and upsample and $\mathbf{W} \in \mathbb{R}^{\frac{HW}{r^2} \times \frac{HW}{r^2}}$ is a linear projection matrix performed on spatial dimension. The LGF module uses global information to estimate the per-location weight for the fused feature $\mathbf{p}+\mathbf{q}$. It then combines the fused feature and decoder features using the predicted weight. The model can therefore learn to disregard unsuitable features from the encoder. Empirically, we set $r = 2^{\log_2 H - 5}$ when $H > 2^5$, otherwise $r = 1$.

3.1.3 Balancing generation and reconstruction

Ideally, a face restoration model should emphasize face generation than reconstruction when there exists severe degradation in the input image and vice versa because a severely

degraded face may not contain sufficient details for reconstruction. Given that a successful face restoration model should handle various types and strengths degradations, it is important to strike a balance between the two sub-tasks. However, our empirical analysis shows that the skip connections in the U-Net architecture impose a strong condition on the generation model and may bias the model toward reconstruction. The more skip connections we add, starting from higher to lower layers, the stronger reconstruction the model performs. See Figure 3.

Previous works [20, 36, 40] choose to apply skip connections in all layers. In contrast, we propose to re-balance the generation and reconstruction sub-task to improve the overall restoration performance. This is achieved by reducing the number of skip connections, particularly skip connections in the lower layers, because low-level skip connections tend to impose stronger conditions on the generation model and weaken its generalization ability. Furthermore, low-level features tend to be less informative in low-quality inputs, given that the degradation may corrupt the information. Empirical results show that this strategy helps to improve face restoration performance. Please refer to the experiments and Appendix for more information.

3.2. Improving Generation

Besides authentic reconstruction, a successful face restoration model also needs to generate realistic high-quality faces. As mentioned before, this is usually achieved through the face generation sub-task encouraged by the adversarial loss. However, empirical results show that prior works that *do not* utilize a pre-trained generator often produce lower quality faces, e.g. DFDNet in Figure 1. In other word, the end-to-end learned generators do not perform as well as off-the-shelf generative networks. To generate crispy and clear faces, we next introduce how to improve the generation sub-task in face restoration.

We hypothesize that the problem of prior works is that they try to learn a deterministic face restoration model G. In contrast, off-the-shelf generative networks are trained non-deterministically by taking random noises as the model inputs. Based on the hypothesis, we propose to learn a stochastic face restoration model by introducing a noise term ϵ,

$$\hat{x} = G(x, \epsilon),\ \epsilon \sim \mathcal{N}(0, 1), \tag{5}$$

motivated by state-of-the-art GAN models [16].

The stochastic model is beneficial from various aspects. It helps to capture the non-deterministic nature of the face restoration problem, where multiple high-quality images may exist that can degrade to the same low-quality face. A deterministic model is unable to capture the desired inverse degradation function. It also helps to better explore the latent feature space for the generation model during training. While a common practice is to sample the degradation function to generate random inputs $x \sim Deg(y)$ during training, we can observe that the input x and the target output y are usually fairly similar, e.g. Figure 1 and Figure 3. As a result, the variations in the latent features may be limited during training, and the generator may not generalize well. By injecting the noises into the latent feature space, the generator may be able to handle more complex cases similar to recent facial prior-based techniques [36, 40].

In practice, we implement the stochastic face restoration model as follows. Let $Enc(x) \in \mathbb{R}^{H' \times W' \times C}$ denote the final feature map extracted by the encoder. We compute the conditional random noises ϵ_c by applying a linear soft gate on ϵ:

$$\epsilon_c = \mathsf{Sigmoid}(z) * \epsilon, \tag{6}$$

where $z = AttentionPool(Enc(x)) \in \mathbb{R}^C$ and $*$ denotes element-wise multiplication [26]. We then feed the noise signals ϵ_c to the decoder, where we implement the decoder based on StyleGAN2 architecture. More specifically, we apply a style-block to both the skip-connection features and decoder features before fusing them using LGF described in Sec. 3.1.2, and we feed ϵ_c to the two blocks by mapping it to the style vector in StyleGAN2. Please see the appendix for implementation details. Compared with unconditional random noises, ϵ_c encapsulates the latent representation z of the input and thus imposes more content-aware control.

3.3. Learning Objective

This section describes the objective function for training. We instantiate the face restoration problem, i.e. Eq. 1, using the following objective function:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{ADV}} + \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{VQ}}. \tag{7}$$

The first two terms are the adversarial generation loss and reconstruction losses and correspond to the two terms in Eq. 1. The last term is the feature quantization loss described in Section 3.1.1. α is a hyper-parameter that balances generation and reconstruction. See appendix for ablation study on the impact of α.

In practice, we implement $\mathcal{L}_{\text{ADV}}$ using non-saturating loss [8] and optimize the model by alternating between optimizing the discriminator D by minimizing

$$-\mathbb{E}_{y \sim Y} \log \left[D(Aug(y))\right] - \mathbb{E}_{x \sim X} \log \left[1 - D(Aug(G(x))\right]$$

and optimizing the generator G by minimizing

$$-\mathbb{E}_{x \sim X} \log \left[D(Aug(G(x))\right],$$

where $Aug(\cdot)$ is the differentiable data augmentation [45] including random color transform and translation. The reconstruction loss is implemented by

$$\mathcal{L}_{\text{REC}} = \mathcal{L}_1 + \mathcal{L}_{percep}, \tag{8}$$

where $\mathcal{L}_1$ is the L1-loss between the target and restored image and $\mathcal{L}_{precep}$ is the perceptual loss based on a pre-trained VGG-19 network [33] following existing works in image generation [7, 13, 20, 36]. See Appendix for details.

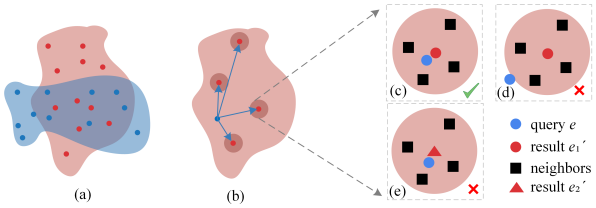


Figure 4. Illustration of iPrecision. (a) Precision measures the portion (overlapped area) of restored images (blue region) that fall into real images (red region). (b) For each restored image, we determine whether it falls into real image manifold by calculating its vectorized feature distance to every real image. (c)-(e) show the decision of one restored image e. We consider four neighbors of each real image and identities satisfy $I_e = I_{e_1'}, I_e \neq I_{e_2'}$.. (c) e is the nearest neighbor of e_1' and both have the same ID. (d) show e is not inside the k-nearest neighborhood. (e) e and e_2' have different IDs though e is the nearest one. Among (c)-(e), only (c) counts as a correct match with iPred $= 1$.

4. Identity Preservation Metric

This section introduces a new evaluation metric that is designed specifically for the face restoration problem. As mentioned previously, prior works in face restoration usually adopt metrics that are intended for general image reconstruction or generation. While these metrics can measure the generic quality of the reconstructed image, they were unable to capture subtle changes on faces that are minor in pixel space but are perceptually significant. In particular, they often fail to measure whether the restored faces preserve the identity-related details.

To address these issues, we propose a metric that simultaneously measures image quality and facial details preservation. The new metric is based on the improved precision and recall metric for the generative model introduced in [19], where precision measures whether the distribution of generated images falls into the distribution of real images and recall measures the opposite. Therefore, a high precision indicates high generated image quality, given that realistic images refer to high-quality images Y in the face restoration problem. We extend the metric to consider facial details preservation, which is measured by the ability to preserve the identity information. In other words, instead of considering whether the restored face falls into the distribution of all high-quality faces, we consider whether it falls into the distribution of the high-quality faces of the same subject. As a result, a high precision implies that the generated faces are high quality and preserve the subject's identity.

More specifically, the evaluation metric is defined as follows. Given a pre-trained feature extractor, e.g. Inception V3 [34] or FaceNet [31], we calculate two sets of image features $\{\mathbf{E}_g, \mathbf{E}_r\}$ that corresponds to the generated and real faces respectively. Let $\tilde{\mathbf{E}}_g{=}\mathbf{E}_r$ and $\tilde{\mathbf{E}}_r{=}\mathbf{E}_g$. For each feature $e \in \mathbf{E}$, we define a binary function

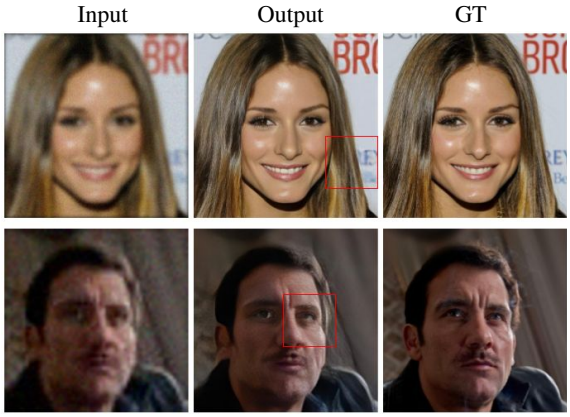


Figure 5. Qualitative examples that illustrate the advantage of the proposed metric. (Top) PSNR=22.8, iPred=1. (Bottom) PSNR=27.5, iPred=0. The top row shows that iPred focuses on the face region and is less sensitive to the artifacts in the back ground. In contrast, PSNR or SSIM place globally equal weight at each pixel. The second row shows that iPred is more sensitive to the artifacts near face components, i.e. the right eye.

$$\text{iPred}(e, \tilde{\mathbf{E}}) = \begin{cases} \mathbf{1}(I_e = I_{e'}), & \exists\, e' \in \tilde{\mathbf{E}} \\ & \text{s.t. } \kappa(e, e') \leq \kappa(e', \text{NN}_k(e', \tilde{\mathbf{E}})) \\ 0, & \text{otherwise} \end{cases}$$

where $I_e, I_{e'}$ are the identity label of e and e' respectively, $\text{NN}_k(e', \mathbf{E})$ is the kth nearest neighbor of e' in $\mathbf{E}$, and $\kappa(\cdot)$ is the Euclidean distance. The binary function indicates whether e falls into the distribution of $\{e'\} \subseteq \tilde{\mathbf{E}}$, where e and e' belongs to the same identity and the distribution of $\{e'\} \subseteq \tilde{\mathbf{E}}$ is represented using the hyperspheres around e'. See Figure 4 for illustration. Given iPred$(\cdot)$, we can define

$$\text{iPrecision}(\mathbf{E}_r, \mathbf{E}_g) = \frac{1}{|\mathbf{E}_g|} \sum_{e_g \in \mathbf{E}_g} \text{iPred}(e_g, \mathbf{E}_r) \quad (9)$$

$$\text{iRecall}(\mathbf{E}_r, \mathbf{E}_g) = \frac{1}{|\mathbf{E}_r|} \sum_{e_r \in \mathbf{E}_r} \text{iPred}(e_r, \mathbf{E}_g) \quad (10)$$

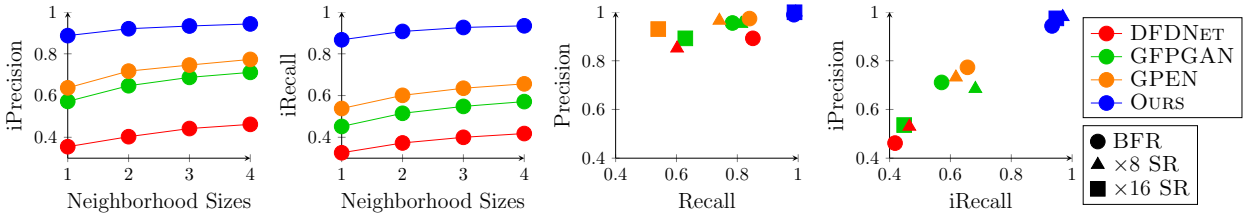
Please refer to the appendix for the pseudo-code. As mentioned before, iPrecision is a good indicator for measuring a face restoration model's actual capability of producing high-fidelity and faithful restorations. This is verified by our user study, which shows that iPrecision better correlates with human evaluations results than standard metrics such as PSNR and LPIPS. Also, see Figure 5 for qualitative examples that illustrate the advantage of the proposed metric.

5. Experiments

We evaluate the performance of the proposed model on standard benchmarks for face restoration. The goal is to verify that 1) the proposed method improves face restoration performance, and 2) the proposed evaluation metric better captures the perceptual image quality in face restoration.

Models	BFR				SR					
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑		LPIPS↓		FID↓	
					×8	×16	×8	×16	×8	×16
DeblurGANv2 [18]	25.91	0.695	0.400	52.69	-	-	-	-	-	-
PSFRGAN [4]	24.71	0.656	0.434	47.59	-	-	-	-	-	-
HiFaceGAN [39]	24.92	0.620	0.477	66.09	26.36	**24.66**	0.211	0.266	29.95	36.26
DFDNet [20]	23.68	0.662	0.434	59.08	25.37	23.11	0.212	0.266	29.97	35.46
mGANprior [9]	24.30	0.676	0.458	82.27	21.44	21.29	0.521	0.518	104.20	100.84
PULSE [23]	-	-	-	-	24.32	22.54	0.421	0.425	65.89	65.33
pSp [29]	-	-	-	-	18.99	18.73	0.415	0.424	40.97	43.37
GFPGAN [36]	25.08	0.678	0.365	42.62	23.80	19.67	0.293	0.382	36.67	63.24
GFPGAN* [36]	24.19	0.681	0.296	38.15	24.12	21.77	0.298	0.342	34.22	37.61
GPEN [40]	23.91	0.686	0.331	25.87	24.97	23.27	0.322	0.361	30.49	31.37
Ours	**28.01**	**0.747**	**0.224**	**18.87**	**26.58**	24.17	**0.205**	**0.260**	**18.27**	**22.94**

Table 1. Quantitative comparison on blind face restoration (BFR) and super-resolution (SR). GFPGAN* denotes the model without colorization. ('-' indicates the number of not available.)



(a) iPrecision and iRecall on BFR (b) Precision vs. Recall on different tasks

Figure 6. Identity-preservation metrics. The color indicates the model, and the marker indicates the task. Best viewed in color.

Training datasets We train our model on the FFHQ dataset [15] and the training split of the CelebA-HQ dataset [14], which consists of 70k and 27k images respectively. All images are resized to 512×512 with Pillow.Image.LANCZOS. Following the standard practice in face restoration [20,21,36,40], we synthesize degraded low-quality faces x from real high-quality faces y using the following degradation model:

$$x = [(y \otimes \mathbf{k}_\sigma)_{\downarrow_r} + \mathbf{n}_\delta]_{JPEG_q}, \tag{11}$$

i.e. the high-quality image y is first convolved with a Gaussian blur kernel $\mathbf{k}_\sigma$ with kernel size σ and downsampled by a factor r. An additive white Gaussian noise $\mathbf{n}_\delta$ with standard deviation δ is then added before applying JPEG compression with quality factor q to obtain the final low-quality image x. The restoration model is trained with image pairs (x, y) following Eq. 7. Please refer to the appendix for implementation details.

The degradation model simulates real world low-quality images caused by defocus, long-distance sensing, noises, compression, and their combinations [21]. While other types of degradation are possible, we adopt the same degradation model used in prior works [20,36,40] for a fair comparison. Similarly, we randomly sample σ, r, δ and q from [0.2, 10], [1, 8], [0, 15] and [60, 100] for the degradation function following GFPGAN [36].

Evaluation metrics We compare our model and the baselines on all 3k images in the test split of CelebA-HQ using two tasks, i.e. Blind Face Restoration (BFR) and Super-Resolution (SR). For BFR, we synthesize low-quality images using the same degradation model as the training data. For SR, we create two sets of low-quality images with resolution 64×64 and 32×32 respectively for ×8 and ×16 SR tasks. We evaluate the performance using 1) standard objective metrics including PSNR, SSIM, LPIPS and FID, 2) the proposed iPrecision and iRecall metrics, and 3) subjective evaluation through user study.

5.1. Objective Evaluation

We first evaluate the model performance using standard objective metrics. Table 1 summarizes the results. Our model consistently outperforms all the baselines with a large margin on both BFR and SR. The results verify that our model exceeds state-of-the-art face restoration models in terms of both the restored image quality and the reconstruction accuracy. The best performing baselines are those that exploit pre-trained StyleGAN generator, i.e. GFPGAN and GPEN. The result shows that a robust image generation model helps to improve the overall restoration performance. Nevertheless, our model outperforms GFPGAN and GPEN while using fewer parameters (50M parameters versus 70M parameters in GPEN and 80M parameters in GFPGAN), which indicates the importance of balancing the generation and reconstruction sub-tasks.

Next, we compare the performance using the proposed identity preservation metrics with FaceNet feature extractor. We focus on comparing with GFPGAN, GPEN, and DFDNet because 1) they achieve the best overall performance among all baselines, and 2) they share the same degradation model with ours during training. The results are in Fig-

Figure 7. Qualitative comparison. (*Top*) BFR. Note the eyelash and skin tone difference. (*Bottom*) $\times 16 : 32^2 \rightarrow 512^2$ SR. Note the expression and wrinkle differences.

Methods	PSNR↑	LPIPS↓	iPrecision↑	Preference (%)↑
Bicubic	<u>26.62</u>	0.361	0.482	0.8
GFPGAN	24.12	<u>0.298</u>	0.687	5.4
GPEN	24.97	0.322	0.732	7.4
Ours	26.58	**0.205**	**0.980**	**86.4**

Table 2. Metric comparison on $\times 8$ SR.

Fusion types	PSNR↑	SSIM↑	LPIPS↓	FID↓
Baseline	26.85	0.710	0.251	20.02
+ LGF	27.13	0.729	0.243	19.55
+ Quantization	27.35	0.737	0.238	19.77
+ Noise	**27.40**	**0.738**	**0.225**	**19.12**

Table 3. Ablation results.

ure 6. Again, our model consistently outperforms the baselines, which shows that our model generates higher quality faces and better preserves the identity-related details in the restored faces. See appendix for results on SR.

Note that the proposed metric has a meta-parameter k, which determines the size of the target distribution. Figure 6(a) shows that, while both the precision and recall improve as k increases, the relative performance of different models remains stable. Therefore, a single k should be sufficient for evaluation, and we set k=4 in the following experiments. Figure 6(b) compares the results of the original precision-recall metrics and the proposed identity-preserving metrics. The results show that the identity information increases the dynamic range of the metrics, which helps to discriminate the performance of different models.

5.2. Subjective Evaluation

We also compare different face restoration models with subjective evaluation. We conducted a user study over 100 randomly selected samples. For each sample, we present the restoration results of four different methods and the input and target images as references to raters. We then ask the rater which image has the best perceptual quality while preserving the facial details in the target image. Five raters annotate each sample, and we measure the percentage of examples where the raters prefer the result of a model. Please refer to the appendix for details.

The user study results on SR are in Table 2. The subjective evaluation again verifies the superior performance of our model. Interestingly, the advantage of our model is much more significant in the subjective evaluation than in objective metrics. This shows that minor changes in the pixel space may significantly impact the perceptual quality in face restoration, and standard metrics like PSNR cannot capture user preference very well. The results also show that the proposed iPrecision metric better correlates with raters' opinion, which justifies the proposed metric's benefit. See appendix for results on BFR.

Figure 7 presents the qualitative examples of different models. The results show that our method can achieve the best perceptual quality and faithfully restore most source details. See appendix for more qualitative results and difficult restoration results.

5.3. Ablation Study

We conduct ablation studies to understand how each model component affects the performance. For fast validation, we apply $1/2$ size of a previously used model. The results are in Table 3, were each of the proposed improvement boost the overall performance. See appendix for details and more ablation results

6. Conclusion

This work revisits the face restoration problem. We show that the face restoration problem can be decomposed into two sub-tasks, i.e. face generation and face reconstruction, and that the issues of existing models stem from the failures in the two sub-tasks. To address the practical problems, we introduce a new model by improving the model design for better generation and reconstruction. We further propose a new objective metric that simultaneously assesses a model's generation and reconstruction performance. Future work will explore personalized face restoration by exploiting additional references or text guidance.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. 11

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 3

[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 *CVPR*, 2018. 1

[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 *CVPR*, 2021. 7

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 1

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 4

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 5

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 5

[9] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 2, 7

[10] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, 2019. 2

[11] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *CVPR Workshops*, 2017. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

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

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 7, 11

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3, 5

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 11

[18] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 2, 7

[19] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019. 6

[20] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020. 1, 2, 4, 5, 7, 17

[21] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 1, 2, 7

[22] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *CVPR*, 2020. 2

[23] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 1, 2, 7

[24] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 4

[25] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 5

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 4

[28] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 4

[29] 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 *CVPR*, 2021. 2, 3, 7

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 3

[31] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 6

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

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 6

[35] Xiaoguang Tu, Jian Zhao, Qiankun Liu, Wenjie Ai, Guodong Guo, Zhifeng Li, Wei Liu, and Jiashi Feng. Joint face image restoration and frontalization for recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1

[36] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 1, 2, 4, 5, 7

[37] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 2

[38] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323*, 2021. 2

[39] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *ACM Multimedia*, 2020. 2, 7

[40] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 1, 2, 4, 5, 7

[41] Rajeev Yasarla, Federico Perazzi, and Vishal M Patel. Deblurring face images using uncertainty guided multi-stream semantic networks. *IEEE Transactions on Image Processing*, 29:6251–6263, 2020. 2

[42] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 2

[43] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 2

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 3

[45] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 5

[46] Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, and Changyou Chen. Feature quantization improves gan training. In *ICML*, 2020. 4

[47] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 2