RETHINKING DEEP FACE RESTORATION

Anonymous authors
Paper under double-blind review

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 facial changes may alter the identity and significantly degrade the perceptual quality. In this work, we argue 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*** by learning a stochastic model and enhancing the latent features respectively. Furthermore, we adapt the number of skip connections for a better balance between the two sub-tasks. 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. The user study shows that our model produces higher quality faces while better preserving the identity 86.4% of the time compared with the best performing baselines.

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 real world due to a variety of reasons: image resolution, motion blur, defocus blur, sensor noises, encoding artifacts and 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 (GAN) (Goodfellow et al., 2014). Existing works treat face restoration as a conditional image generation problem and often resort to U-Net architectures to restore high-quality faces from low-quality images (Bulat & Tzimiropoulos, 2018; Li et al., 2018; 2020; Menon et al., 2020; Yang et al., 2021; Wang et al., 2021). 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. For example, the model may change the subject's eye color or change the eyelids from monolid to double eyelid, as shown in Figure 1. Usually these changes are negligible in pixel space and irrelevant for realisticness but essential for authenticity, which can lead to biometric characteristics deviate from the original subject, thus may significantly degrade the perceptual quality, especially for people familiar with the subject.

In this paper, we argue that the above issues are caused by the fragile balance between face generation and face restoration. We show that the face restoration problem is 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 (Choi et al., 2020; Wang et al., 2021). 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 lead to unsatisfactory restoration quality. Therefore, a successful face restoration model has to adequately address the two sub-tasks simultaneously, which remains to be realized.

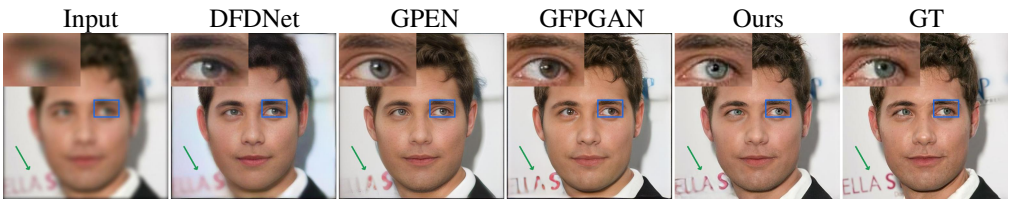

Figure 1: Problems of state-of-the-art face restoration models. GPEN and GFPGAN are biased toward face generation and may alter facial details (e.g. eye color) that are highly correlated with the identity. DFDNet is biased toward reconstruction and does not remove all degradations. Our approach achieves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

Therefore we propose a new model that aims to improve both generation and reconstruction. To 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 of the U-Net architecture 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 architecture of the model, particularly the number of skip connections, to optimize the balance between generation and reconstruction.

Similar to the approaches, 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) (Heusel et al., 2017), 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) (Zhang et al., 2018b). 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 that may alter the identity. To this end, we propose a new evaluation 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 the best performing baselines.

2 Related Work

Face restoration Face image restoration has attracted considerable attention with a wide range of valuable topics, *e.g.,* face super resolution (Guo et al., 2017; Wang et al., 2018; Menon et al., 2020; Yang et al., 2020; Ma et al., 2020), blind face restoration (Li et al., 2018; 2020; Wang et al., 2021; Yang et al., 2021), deblurring (Yasarla et al., 2020; Shen et al., 2018; Kupyn et al., 2019), denoising (Zhang et al., 2018a; Guo et al., 2019), inpainting (Yu et al., 2018; Zheng et al., 2019), etc. Unlike other image domains, human perceptions are more sensitive to facial images 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 (MLE) to recover realistic face characteristics and adversarial learning to generate a high-fidelity image distribution.

Current state-of-the-art BFR models finetune a GAN-prior network (Gu et al., 2020; Richardson et al., 2021; Wang et al., 2021; Yang et al., 2021). This line of works takes advantage of the fact that most high-dimensional data have support in the lower-dimensional manifold. Therefore, in most cases, we can expect the prior network produces high-fidelity faces by mapping the degraded faces into this lower-dimensional latent space. Proper finetuning, including iterative optimization (Menon et al., 2020) at inference and model improvements (Wang et al., 2021; Yang et al., 2021), further

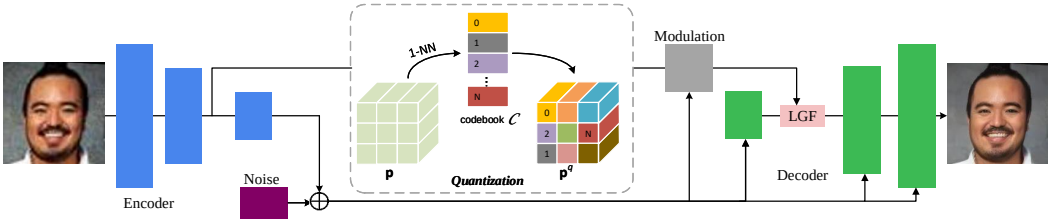

Figure 2: The proposed model with one skip connection. (1-NN: 1-nearest neighbor search. Modulation: feature modulation as in StyleGAN2 (Karras et al., 2020). LGF: linear gated feature fusion.)

enhances the identity information. Although they have shown inspiring performances, the prior knowledge still dominates throughout the finetuning process, leading to unfaithful restoration, *e.g.,* color shift and excessive image completion. In other words, **when the generator has enough capacity, high-fidelity generation is easy, however, faithful generation is hard.** Notably, our approach can be trained from scratch and reaches a good balance between content preservation and generalization ability after tracing the causes of the above issues.

Evaluation metric Current works adopt PSNR, SSIM, and LPIPS to measure the restoration quality for every example while using FID to evaluate how the restored image distribution approaches ground-truth distribution. However, they may cause inconsistent judgment from one to another. A well-known example is blurring images that have better PSNR and SSIM (Zhang et al., 2018b). FID is affected mainly by the number of evaluation samples and may also bring unfair comparisons without prior knowledge of the evaluation system (Parmar et al., 2021). 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 at the distribution level and individual sample's 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 problem of face restoration and then introduce how to improve the generation and reconstruction sub-tasks. Finally, we describe the objective function for training.

Let X denote the degraded low-quality image domain, Y denote the high-quality image domain, and P_Y denote 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 a inverse function G: $X \rightarrow Y$ that satisfies

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

where $\mathcal{D}$ is a distribution distance measure and $\kappa(\cdot)$ is a pair-wise distance between two images. The first half of Eq. 1 encourages the restored images to look realistic and be indistinguishable from real high-quality images. The second half of Eq. 1 encourages the restored image to preserve facial features in the high-quality image from which the input image is degraded from. A common practice is to implement G based on U-Net architecture as illustrated in Figure 2 and implement the first and second half of Eq. 1 using an adversarial loss and reconstruction losses respectively.

Eq. 1 clearly shows that the face restoration problem is a combination of the face generation and face reconstruction sub-task. The generation sub-task is driven by $\mathcal{D}(P_{G(X)}||P_Y)$ and aims to learn the distribution of real high-quality image. It can be further mapped to the decoder in G, which learns to generate realistic image from a latent feature. On the other hand, the reconstruction sub-task is driven by $\mathbb{E}_{y\sim Y}\mathbb{E}_{x\sim Deg(y)}\kappa(G(x), y)$. It aims to learn a feature extractor that projects an image to the latent feature space of the generation model such that the corresponding high-quality image may be generated from the extracted feature. To restore images with different degradations, the feature extractor also has to be robust to the degradation in the input image. Based on this interpretation, we next describe how to improve the generation and reconstruction sub-tasks respectively to achieve better face restoration.

3.1 Improving Reconstruction

As mentioned before, the face reconstruction sub-task requires fine-grained control on face details in the generated image based on the input image. 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 passes low to high level information to the decoder for an authentic reconstruction of the input face.

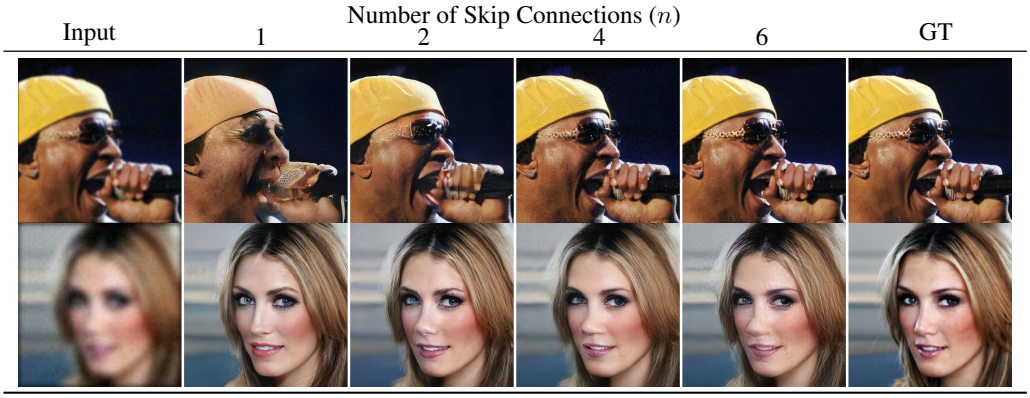

Figure 3: Qualitative comparison by varying the number of skip connections. We count from the layer with feature resolution 8×8, *i.e.,* there exist possible skip connections at resolution nodes $\{2^{n+2} \times 2^{n+2}\}_{n=1}^{6}$ when we set the maximum input resolution at 512×512.

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 were unable to extract useful low-level features from low quality images, and the low quality features hinder the restoration performance. To address this issue, we propose the following improvements for the U-Net architecture.

3.1.1 Balancing Generation and Reconstruction

Ideally, a face restoration model should put more emphasis on face generation than on 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 degradations of various types and strength, it is important to strike a balance between the two sub-tasks. However, our empirical analysis show that the skip connections in the U-Net architecture imposes 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.

To improve the overall restoration performance, we propose to re-balance the generation and reconstruction sub-task. This is achieved by reducing the number of skip connections, particularly skip connection 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 information may be corrupted by the degradation. Empirical results show that this strategy help to improve the face restoration performance. Please refer to the experiments and Appendix C.1 for more information.

3.1.2 Feature Quantization

To help the model generalize to severely degraded image, 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 (Oord et al., 2017; Razavi et al., 2019; Zhao et al., 2020b; Esser et al., 2021; Ramesh et al., 2021) 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 word, 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 vector 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 by optimizing

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

where $\text{sg}(\cdot)$ is a stop-gradient operator. Note that only the first term in $\mathcal{L}_{\text{VQ}}$ optimizes the codebook while the second term encourages the model to utilize quantized feature. In practice, we approximate the first term using exponential moving average (EMA) and optimize the model using the second term only, following prior works on feature quantization (Oord et al., 2017; Razavi et al., 2019). To ensure that the codebook contain only high-quality features and contain useful information for reconstruction, EMA is computed over features extracted from ground truth high-quality images.

3.1.3 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 (Yang et al., 2021), or spatial feature transform (Li et al., 2020; Wang et al., 2021) 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 filters 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 $\mathbf{p} + \mathbf{q}$ and then combine 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 $r = 1$.

3.2 Improving Generation

As reconstruction-based restoration often produces blurry faces, we further resort to adversarial learning to generate crispy and clear faces, as in Eq. 1. Most face restoration approaches (Li et al., 2020; Wang et al., 2021; Yang et al., 2021) treat G as a deterministic function where each input x is associated by only one output $\hat{x} = G(x)$. In most scenarios, we can observe that the input x and the output $\hat{x}$ are usually not far away from each other, *e.g.* Figure 1 and Figure 3. This peculiarity will lead x to become a strong conditional signal where $G(x)$ largely relies on deep internal features of x, *e.g.,* textures and shapes. The internal skip connections can further intensify those signals. However, as the real degradation functions $Deg(\cdot)$ is usually unknown, strong conditions usually fail representation learning and prevent the model's generalization ability, *e.g.,* the second row in Figure 3. We propose a stochastic restoration model to increase the generation power.

3.2.1 From deterministic to stochastic

We claim that it is beneficial to assume G as a stochastic function by introducing a noise term ϵ,

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

Gaussian noise has a relatively high bandwidth to deal with various degradation scenarios by confining high-dimensional data into a low-dimensional manifold. It is consistent with the intuition that recent facial prior-based techniques can handle more complex cases than training from scratch (Wang et al., 2021; Yang et al., 2021). We herein propose a generic approach by perturbing the correlated low-quality skip features with independent Gaussian noises.

As shown in Figure 2, we connect the noise signals to two parts: *decoder blocks* and *skip connection blocks*. Using noise in the decoder part as a modulation signal has been seen in StyleGAN-related literature (Karras et al., 2019; 2020; Choi et al., 2020; Zhao et al., 2021). The purpose of injecting noise in skip-connection is somewhat different. As we claimed in Section 3.1, skip connections are crucial to maintaining source contents.

Adaptive latent gate Let Enc denote the encoder. Assume the latent vector $z = Enc(x)$, we enable conditional noises ϵ_c by applying a linear soft gate on ϵ:

$$\epsilon_c = \mathsf{Sigmoid}(z) * \epsilon, \epsilon \sim \mathcal{N}(0, 1) \tag{6}$$

where $*$ denotes element-wise multiplication. The formulation intermediately yields two advantages, specifically for face restoration. Firstly, ϵ_c encapsulates the input representation z and thus imposes more content-aware control on the multi-scale features than unconditional random noises. Secondly, in practice, ϵ_c is a scaled version of ϵ, which is gradually learned to *implicitly* control a single sample's quality by reducing overall samples' variety (Brock et al., 2018; Karras et al., 2019). Consequently, both coincide with the goal of content preservation from the perspective of a generative model.

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 3.1.2. 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 (Goodfellow et al., 2014) and optimize the model by alternating between Optimize D:

$$\min_{D} \; -\mathbb{E}_{y \sim Y} \log \left[D(Aug(y)) \right] - \mathbb{E}_{x \sim X} \log \left[1 - D(Aug(G(x)) \right] \tag{8}$$

Optimize G (partially):

$$\min_{G} \; -\mathbb{E}_{x \sim X} \log \left[D(Aug(G(x)) \right], \tag{9}$$

where D is the discriminator and $Aug(\cdot)$ is the differentiable data augmentation (Zhao et al., 2020a) including random color transform and translation. The reconstruction loss is implemented by

$$\mathcal{L}_{\text{REC}} = \mathcal{L}_1 + \mathcal{L}_{percep}, \tag{10}$$

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 (Simonyan & Zisserman, 2014) following existing works in image generation (Gatys et al., 2016; Johnson et al., 2016; Wang et al., 2021; Li et al., 2020). See Appendix A for details.

4 A Metric for identity preservation and perceptual quality

Current face restoration evaluation system, *e.g.*, PSNR, FID and LPIPS, is not designed for the purpose of face restoration. Specifically, PSNR is not sensitive to perceptual quality, FID doesn't consider pairwise content preservation and LPIPS excludes face-specific identities.

Inspired by the proposed precision and recall metric in (Kynkäänniemi et al., 2019), we adapt it to the face restoration tasks to simultaneously measure both perceptual quality and identity preservation. We present two metrics iPrecision and iRecall both of which measure the probability of one distribution falling into another distribution by considering image identities. Concretely, iPrecision measures the probability of generated images overlaps with real images, and iRecall measures the probability of real images overlaps with generated images. Figure 4 illustrates the idea of identity-preserved iPrecision. The two metrics naturally conclude the perceptual distance in the feature space such that it can indicate the perceptual quality. By adding identity information, it is more beneficial to evaluate face restoration. Therefore, it is worth noting that the two measured aspects coincide with the goals of *generation* and *reconstruction* respectively. The metric evaluation process includes two steps:

Feature prediction Given a pretrained feature extractor, *e.g.,* Inception V3 (Szegedy et al., 2015) or FaceNet (Schroff et al., 2015), we calculate two sets of image embeddings as $\{\mathbf{E}_g, \mathbf{E}_r\}$, respectively corresponding to the paired restored images and real images. For each feature e, we use a face

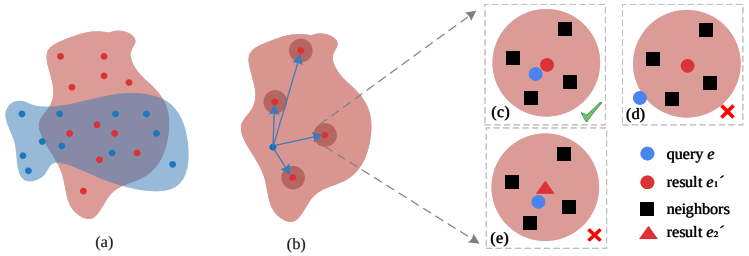

Figure 4: Illustration of iPrecision. (a) Precision measures the portion (overlapped area) of restored images (blue region) that fall into real images (red region). (b) For each restored image, we determine whether it falls into real image manifold by calculating its vectorized feature distance to every 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 $\mathsf{iPred} = 1$.

identity-related binary prediction $\mathsf{iPred}(\cdot)$ to get a relative prediction in the disjoint set $\mathbf{E}$, *e.g.*, $\{\mathsf{iPred}(e_g, \mathbf{E}_r)|e_g \in \mathbf{E}_g\}$ and $\{\mathsf{iPred}(e_r, \mathbf{E}_g)|e_r \in \mathbf{E}_r\}$,

$$\mathsf{iPred}(e, \mathbf{E}) = \begin{cases} \mathbf{1}(I_e = I_{e'}), & \exists\, \kappa(e, e') \leq \kappa(e', \mathrm{NN}_k(e', \mathbf{E})) \\ 0, & \text{otherwise} \end{cases}$$

where $I_e, I_{e'}$ are face identities and $\mathrm{NN}_k(\cdot)$ returns the kth nearest feature by querying the feature e to the set $\mathbf{E}$. We choose Euclidean distance function as $\kappa(\cdot)$. This prediction takes both feature-level similarity and real face identities into consideration. More importantly, the whole real image set is included to measure how realistic the restored image is (Kynkäänniemi et al., 2019).

iPrecision and iRecall Next, we would like to compute identity-related precision and recall as:

$$\mathsf{iPrecision}(\mathbf{E}_r, \mathbf{E}_g) = \frac{1}{|\mathbf{E}_g|} \sum_{e_g \in \mathbf{E}_g} \mathsf{iPred}(e_g, \mathbf{E}_r) \tag{11}$$

$$\mathsf{iRecall}(\mathbf{E}_r, \mathbf{E}_g) = \frac{1}{|\mathbf{E}_r|} \sum_{e_r \in \mathbf{E}_r} \mathsf{iPred}(e_r, \mathbf{E}_g) \tag{12}$$

Therefore, iPrecision is a good indicator for measuring a face restoration model's actual capability of producing high-fidelity and faithful restorations. In Figure 4, we illustrate the meaning of the proposed metric. We list the pseudo-code for calculating precision in Algorithm 1 in Appendix .

Figure 5 visualizes the advantage of using the proposed metric to measure a face restoration's performance, compared with traditional PSNR and SSIM. The top row shows that PSNR and SSIM somehow fail to detect facial image artifacts though their values are relatively high. The bottom row indicates they also place globally equal weight at each pixel, which is not as desired in face restoration. Because the restored face areas should be relatively more noteworthy than the background. In contrast, the proposed identity-related metric doesn't suffer from these issues. Later, we present quantitative numbers that indicate that the metric correlates better with human evaluations than other metrics.

Figure 5: The advantage of using the proposed metric iPred.

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.

5.1 Experimental Settings

Our model is trained on the full 70K FFHQ and 27K CelebA-HQ training split. The remaining 3K CelebA-HQ images are used for evaluation. In all experiments, images are resized to 512×512

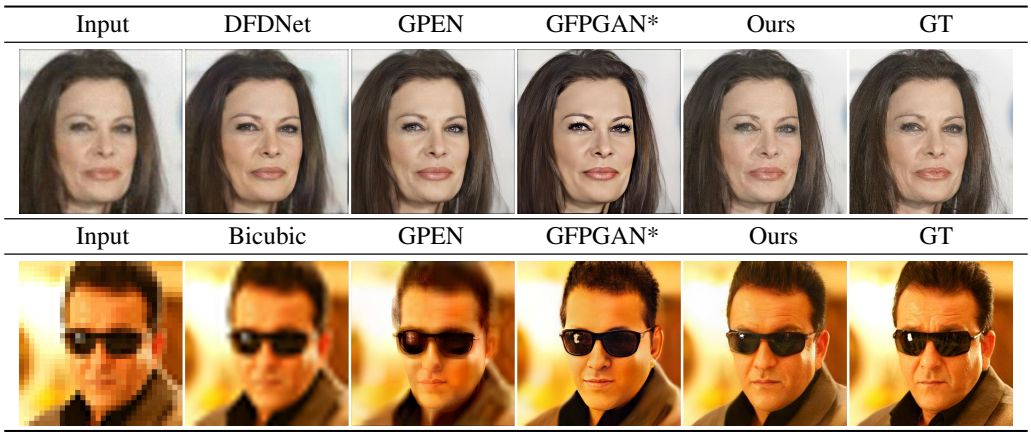

Figure 6: Qualitative comparison. (*Top*) BFR. Note the eyelash and skin tone difference. (*Bottom*) $\times 16 : 32^2 \rightarrow 512^2$ SR. Note the expression and wrinkle differences.

with Pillow.Image.LANZCOS filter. Following previous works (Wang et al., 2021; Li et al., 2020), the training samples x are generated from high-quality face images y from the training set using a degradation function,

$$x = [(y \otimes \mathbf{k}_\sigma)_{\downarrow_r} + \mathbf{n}_\delta]_{JPEG_q} \tag{13}$$

where $\mathbf{k}_\sigma$ is the blur kernel with kernel size σ, r denotes the downsample size, $\mathbf{n}_\delta$ denotes Gaussian noise with standard deviation δ, and q is the JPEG compression ratio. We thus can construct (input, target) image pairs (x, y) and train the model following Eq. 7. Similar to GFPGAN (Wang et al., 2021), we randomly sample σ, r, δ and q from [0.2, 10], [1, 8], [0, 15] and [60, 100]. We evaluate the model performance using (1) standard evaluation metrics including PSNR, SSIM, LPIPS and FID, (2) the proposed iPrecision and iRecall metrics, and (3) user study. See Appendix A for details.

5.2 Objective Evaluation

We compare our model with state-of-the-art approaches on the tasks of Blind Face Restoration (BFR) and Super Resolution (SR). Table 1 shows the comparison across state-of-the-art models on BFR task. Our model achieves the best quantitative numbers on all metrics by a large margin, meaning that our model exceeds all baselines in both image fidelity and content preservation. Figure 6 shows the qualitative results. We can observe that our model can synthesize realistic faces with source contents better preserved. More examples including real world low-quality image restoration samples are given in Figure 12 and Figure 13 in Appendix.

For super resolution, we create two sets of evaluation images with resolution 64×64 and 32×32 respectively for $\times 8$ and $\times 16$ SR tasks. For fair comparison, the resizing method follows original implementation of each approach. As in Table 1, our model achieves the best quantitative numbers on most metrics. Figure 6 shows that our method can achieve the best perceptual performance and faithfully restore most source details. More examples are given in Figure 14 and 15.

Models	**BFR**				**SR**					
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑		LPIPS↓		FID↓	
					×8	×16	×8	×16	×8	×16
DeblurGANv2 (Kupyn et al., 2019)	25.91	0.695	0.400	52.69	-	-	-	-	-	-
PSFRGAN (Chen et al., 2021)	24.71	0.656	0.434	47.59	-	-	-	-	-	-
HiFaceGAN (Yang et al., 2020)	24.92	0.620	0.477	66.09	26.36	**24.66**	0.211	0.266	29.95	36.26
DFDNet (Li et al., 2020)	23.68	0.662	0.434	59.08	25.37	23.11	0.212	0.266	29.97	35.46
mGANprior (Gu et al., 2020)	24.30	0.676	0.458	82.27	21.44	21.29	0.521	0.518	104.20	100.84
PULSE (Menon et al., 2020)	-	-	-	-	24.32	22.54	0.421	0.425	65.89	65.33
pSp (Richardson et al., 2021)	-	-	-	-	18.99	18.73	0.415	0.424	40.97	43.37
GFPGAN (Wang et al., 2021)	25.08	0.678	0.365	42.62	23.80	19.67	0.293	0.382	36.67	63.24
GFPGAN* (Wang et al., 2021)	24.19	0.681	0.296	38.15	24.12	21.77	0.298	0.342	34.22	37.61
GPEN (Yang et al., 2021)	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.)

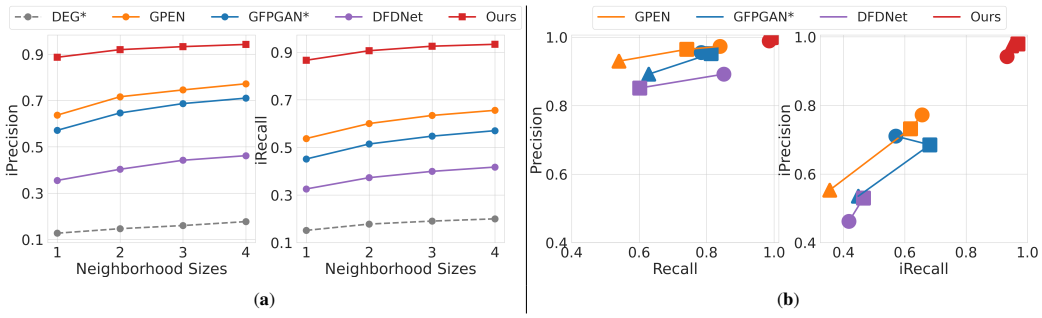

Figure 7: (**a**) iPrecision and iRecall with different neighborhood sizes on BFR. DEG* indicates the created degraded images. (**b**) Precision versus recall from various approaches and different tasks. (◯ denotes BFR task, □ for $\times 8$ SR and △ for $\times 16$ SR.)

5.3 Proposed iPrecision and iRecall metrics

To validate whether the proposed metric is more effective than others in face restoration, we start by ablating neighborhood size as in Figure 7(a), (i) Increasing the neighborhood sizes leads to higher precision and recall by allowing more misses. (ii) Our approach consistently gives the best restoration quality even when we set the neighborhood size $k = 1$, meaning that restored faces with our approach are the closest ones among all 3K testing images to the sources. (iii) Varying neighborhood size would not change ranking order of different methods, demonstrating the robustness of proposed metrics, therefore in our experiments, we set $k = 4$. Moreover, as is shown in Figure 7(b), including hard-coded identity information indeed produces more discriminative numbers than calculating distances only. In a word, low precision and recall tell us that the model is very likely to generate a "fake face" of some different person even if the appearance is sharp. By default, we use FaceNet as the feature extractor and find Inception V3 gives similar result in Appendix B.

5.4 Subjective evaluation

We further conduct a user study to assess the correlation between the proposed metric and human opinions. As is seen in Table 2, our model has achieved the best result and the proposed iPrecision has a better correlation with human opinions. The underlined numbers mean that the metric is inconsistent with human rates. More human study details are provided in Appendix D.

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.

5.5 Ablation study

We conduct ablation studies to understand how each model component affects face restoration performance. For fast validation, we apply $1/2$ size of a previously used model. We here study the impact of the proposed three techniques: linear gated feature fusion, feature quantization and noise injection. As is observed in Table 3, they all can boost the overall performance. In practice, we also find that linear gated fusion is more stable than the other two fusion methods when we increase the degradation level in training. More ablation results can be found in Appendix C.

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 face restoration 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

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. URL https://www.tensorflow.org/.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

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*, pp. 109–117, 2018.

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*, pp. 11896–11905, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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*, pp. 3012–3021, 2020.

Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1712–1722, 2019.

Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 104–113, 2017.

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.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

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*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

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

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8878–8887, 2019.

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

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)*, pp. 272–289, 2018.

Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pp. 399–415. Springer, 2020.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5569–5578, 2020.

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*, pp. 2437–2445, 2020.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

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

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.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pp. 14866–14876, 2019.

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*, pp. 2287–2296, 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

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*, pp. 8260–8269, 2018.

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

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

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*, pp. 9168–9178, 2021.

Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hiface-gan: Face renovation via collaborative suppression and replenishment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1551–1560, 2020.

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*, pp. 672–681, 2021.

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.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

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, 2018a.

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*, pp. 586–595, 2018b.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv:2006.10738*, 2020a.

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.

Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, and Changyou Chen. Feature quantization improves gan training. In *International Conference on Machine Learning*, pp. 11376–11386. PMLR, 2020b.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1438–1447, 2019.