Analyzing and Combating Attribute Bias for Face Restoration

Zelin Li[1,2] , **Dan Zeng**[1,2*] , **Xiao Yan**[2] , **Qiaomu Shen**[1,2] , **Bo Tang**[1,2]
[1]Research Institute of Trustworthy Autonomous Systems,
Southern University of Science and Technology
[2]Department of Computer Science and Engineering,
Southern University of Science and Technology
{lizl2017@mail., zengd@, yanx@,shenqm@,tangb3@}sustech.edu.cn

Abstract

Face restoration (FR) recovers high resolution (HR) faces from low resolution (LR) faces and is challenging due to its ill-posed nature. With years of development, existing methods can produce quality HR faces with realistic details. However, we observe that key facial attributes (e.g., age and gender) of the restored faces could be dramatically different from the LR faces and call this phenomenon *attribute bias*, which is fatal when using FR for applications such as surveillance and security. Thus, we argue that FR should consider not only image quality as in existing works but also attribute bias. To this end, we thoroughly analyze attribute bias with extensive experiments and find that two major causes are the lack of attribute information in LR faces and bias in the training data. Moreover, we propose the *DebiasFR* framework to produce HR faces with high image quality and accurate facial attributes. The key design is to explicitly model the facial attributes, which also allows to adjust facial attributes for the output HR faces. Experiment results show that DebiasFR has comparable image quality but significantly smaller attribute bias when compared with state-of-the-art FR methods.

1 Introduction

Face restoration (FR) is the task of recovering high-resolution (HR) faces from their low-resolution (LR) counterparts and finds many applications such as video surveillance, portal control, and traffic monitoring. FR is challenging as the problem is under-constrained and LR faces can come with severe degradation (e.g., motion blur [Kupyn *et al.*, 2018], noise [Zhang *et al.*, 2017], and JPEG compression [Dong *et al.*, 2015]). Early methods [Zhou *et al.*, 2015; Huang and Liu, 2016] treat face image as natural image. They design methods without exploiting any face characteristics and the output HR faces suffer from the severe over-smooth problem. Latter methods [Song *et al.*, 2017; Chen *et al.*, 2018; Grm *et al.*, 2019] incorporate various face priors (e.g., facial landmarks, parsing maps, and identity information) to improve image quality, but their outputs are still over-smooth and the magnification factor usually does not exceed 8x. State-of-the-art methods, VQFR [Gu *et al.*, 2022], GFPGAN [Wang *et al.*, 2021] and GPEN [Yang *et al.*, 2021] for example, use reference prior or pre-trained generative adversarial network (GAN) priors and produce faces with realistic and clear details as illustrated in Figure 1.

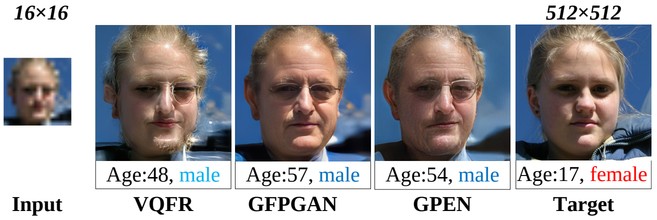


Figure 1: An example of attribute bias, best viewed when zoomed in. The low-resolution face is downsampled from a high-resolution face, and three existing face restoration methods produce realistic output faces. However, key face attributes of the restored faces are dramatically different from the ground-truth high-resolution face.

However, the restored faces produced by state-of-the-art methods can be unreliable as their key facial attributes (e.g., age and gender are altered) can be dramatically different from the ground truth. We show such an example in Figure 1, where a teenage girl is restored to a middle-aged man by three methods. We call this phenomenon *attribute bias*, which can be fatal for applications where facial attributes count, e.g., identifying and tracking criminals for surveillance. We quantify attribute bias as how far the restored faces deviate from the ground truth in attribute and analyze it with extensive experiments, which lead to two main findings. ❶ Attribute bias becomes more severe when there is less attribute information in the input LR faces. In particular, the confidence of pre-trained classifiers for attribute information decreases with the resolution of the LR faces but the attribute bias increases. ❷ Data prior affects attribute bias. Specifically, we construct datasets with different attribute distributions for model training and find that the attributes of the restored faces are biased toward the majority.

Our findings suggest that attribute bias is difficult to tackle

[*]Corresponding author.

§ Code and supplementary materials are available at the link: https://github.com/Seeyn/DebiasFR.

using only the LR faces: low resolution is why we need FR in the first place and it is difficult to collect large datasets with balanced distribution on the attribute. Thus, we consider cases where additional attribute information is provided as input. For instance, in criminal investigation, attribute information of suspects can be obtained from the witnesses, and when restoring old films, attribute information of the characters and actors is available. We aim to produce faces with not only high quality but also small attribute bias. The task is challenging as images and attributes lie in different domains and they need to be coherent for high image quality.

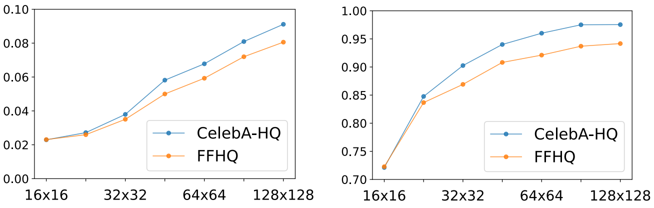
We propose the *DebiasFR* framework for such use cases. DebiasFR explicitly models facial attributes in the latent space of GAN priors that are used to produce the HR face via the decoder model and represents each value in the domain of an attribute as an embedding vector. Thus, the output face can be configured to have certain attributes by activating the corresponding embedding vectors in the latent space. To train attribute embedding for encoding facial attributes, besides the usual pixel-wise restoration loss, we introduce an *attribute loss*, which measures how well the attributes of the output face match the input specifications. As a pair of HR and LR faces allow only one set of attribute information, we propose a *pseudo pair strategy* for data augmentation, which allows using different attribute information for one LR face. For pseudo pairs, we compute the restoration loss by degrading the HR faces into LR ones to tackle the absence of ground-truth HR faces. We compare DebiasFR with state-of-the-art FR methods on three datasets from different scenarios. The results show that DebiasFR matches the baselines in image quality and effectively retains the input facial attributes in the HR faces. Besides manual attribute inputs, DebiasFR can also take attribute information produced by pre-trained classifiers and allows applying different attribute information on the same LR face for trials.

To sum up, we made the following contributions:

- We observe attribute bias in face restoration, which is fatal for many applications and suggests that focusing only on image quality may be problematic.
- We thoroughly analyze attribute bias and trace it to two main causes, i.e., the lack of attribute information and bias in training data.
- We propose DebiasFR, which faithfully preserves input attribute information and produces quality HR faces.

2 Analyzing the Attribute Bias

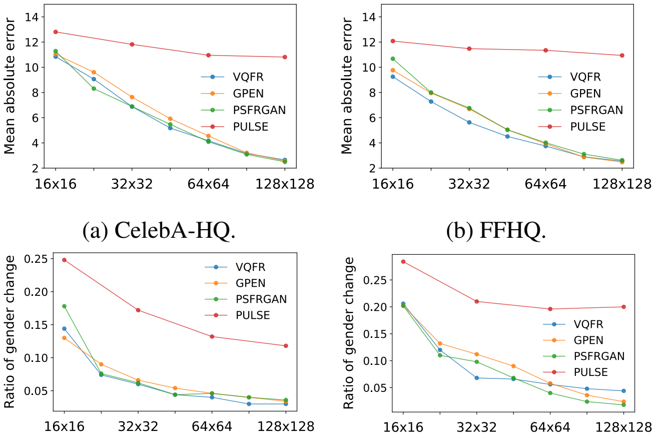
In this section, we systematically analyze the cause of the attribute bias problem. We target widely used face attributes including gender and age for our observation, as they can be easily captured in the real scene and are important biometrics [Hassan *et al.*, 2021]. Meanwhile, our observation experiment is more of a paradigm, it can be extended to include more attributes if there are any. For data selection, we use FFHQ [Karras *et al.*, 2019] and CelebA-HQ [Karras *et al.*, 2018] and their predicted attributes as target images and attributes. Specifically, we use the attribute estimator [Rothe *et al.*, 2018] to generate gender (i.e., belongs to Male, Female) and age (i.e., ranging from 0 to 100 years old). To ensure the universality of the experiment, we conduct experiments with four representative face restoration methods, namely, PLUSE [Menon *et al.*, 2020], PSFRGAN [Chen *et al.*, 2021], GPEN [Yang *et al.*, 2021], and VQFR [Gu *et al.*, 2022].



(a) Age Confidence.

(b) Gender Confidence.

Figure 2: Average attribute prediction confidence for images at different resolutions.



(a) CelebA-HQ.

(b) FFHQ.

(c) CelebA-HQ.

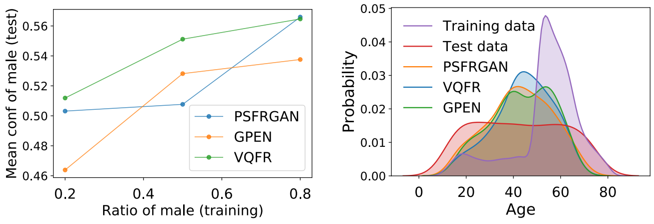
(d) FFHQ.

Figure 3: Average attribute bias of the methods when restoring images of different resolutions.

Observation 1: *The attribute bias increases dramatically as the input image information decreases.*

As illustrated in Figure 2(a) and Figure 2(b), the confidence of both gender and age continues to decrease as the input image resolution decreases. This indicates that the attribute information loses as the image resolution becomes lower. As shown in Figure 3, the lack of attribute information results in an obvious attribute bias enlargement, which is quantified by the mean absolute error of age and the ratio of gender change. As illustrated in Figure 3(a) and Figure 3(b), we can observe an increase in age bias, as indicated by the rise in the mean absolute error. It reaches 10 years when the input resolution decreases to 16 × 16. This phenomenon also occurs with gender, as shown in Figure 3(c) and Figure 3(d), where the ratio of gender change dramatically increases as the resolution decrement.

Experiment settings: we divide the gender attribute into male and female groups, and the age attribute into five groups: 10-20, 20-30, 30-40, 40-50, 50+. We sample 500 images from FFHQ and CelebA-HQ and guarantee an equal num-



(a) Gender.

(b) Age.

Figure 4: Observing the impact of training priors on attribute bias. (a) Impact of the gender ratio and (b) Effect of the age ratio on the attribute bias.

ber of images from each attribute group. The images are then downsampled to several different resolutions including 16^2, 24^2, 32^2, 48^2, 64^2, 96^2, and 128^2 via bilinear interpolation to simulate the loss of image information at different extend.

Observation 2: *The attributes of the recovered images are greatly affected by the attribute distribution of training data.*

As illustrated in Figure 4(a), as the ratio of males in the training data increases from 0.2 to 0.8, we observe the same trend in the confidence of males in the recovered images from 0.46 to 0.56. Specifically, when the ratio of training males is 0.5, that is, half male training images and half female training images, the confidence of males in the recovered images is about 0.5. This finding shows that the recovered attribute information is strongly affected by the attribute distribution of training data, which has not been explored in existing face restoration methods. The same observation is also found in the age attribute in Figure 4(b). Although the age distribution in the test dataset is uniform, different face restoration methods are prone to recovery images with a certain similarity to the age distribution of the training data. For example, all methods tend to restore age attributes to 40-50, which is biased toward the training data distribution. Some qualitative results are included in our supplementary material.

Experiment settings: FFHQ is used as training data and CelebA-HQ is used for testing. We sample 260 images from CelebA-HQ to ensure an equal number of images for each attribute group. All images are downsampled to 32×32 for restoration. For the training dataset, we adjust the attribute distribution to suit our purpose. Specifically, we sample 41350 images from FFHQ to generate subsets with different male ratios, including 0.2, 0.5, and 0.8. Similarly, we sample 14300 images from FFHQ and set the ratio of images for 10-20, 20-30, 30-40, 40-50, and 50+ age groups to 1:1:1:1:6 and use kernel density estimation to estimate the age distribution of training and test sets.

3 The DebiasFR Framework

As depicted in Figure 5, the model consists of three parts: An image encoder, a decoder and some attribute representations. The encoder adopts a U-Net architecture, while the decoder is a fixed pre-trained StyleGAN [Karras *et al.*, 2019]. The attribute representations are embedding vectors in the latent space of the StyleGAN.

The image information is extracted by the encoder and inputted to the decoder by skip-connection structures and latent space in the form of a latent vector. The base latent vector from the encoder is adjusted by the addition of attribute representations which can be viewed as a movement in the latent space. After the addition, the decoder reconstructs the image according to the input from skip-connection structures and latent space. Since the addition of the attribute is independent of the image input, it can be adjusted after the restoration. This process can be formulated as follows:

$$\hat{y} = f_{x,\theta}(\overrightarrow{n} + \lambda\overrightarrow{r}), \tag{1}$$

where x is the input image and $\hat{y}$ is the restored image. θ is the model parameters and f denotes the model. $\overrightarrow{n}$ is the base latent vector extracted by the encoder and $\overrightarrow{r}$ is the attribute representations. λ is the weight of the attribute representations. An adaptive feature fusion module employing the attention mechanism is adopted to better utilize the GAN prior, whose detailed introduction is in supplementary materials.

3.1 Attribute Representation

It has been observed that the latent space of a well-trained GAN model has great properties in attribute manipulation. We find that with the architecture modification, the model still keeps some useful properties from the GAN model.

Property 1: *Given the base latent vector $\overrightarrow{n}$, the attribute of the restored image can be manipulated by moving the latent vector in an interpretable direction.*

The face attribute space is well-bridged with the latent space [Shen *et al.*, 2020; Shen and Zhou, 2021; Härkönen *et al.*, 2020]. If there exists a scoring function s for a face attribute and corresponding direction $\overrightarrow{r}$ for the attribute, then:

$$s(f_{x,\theta}(\overrightarrow{n} + \lambda\overrightarrow{r})) = s(f_{x,\theta}(\overrightarrow{n})) + \lambda\epsilon, \tag{2}$$

with the $\epsilon > 0$. We call the direction an interpretable direction, and this property indicates that fine-grained manipulation can be achieved by finding out the interpretable directions and adjusting the distance. We train attribute representations, which essentially search the direction for corresponding attributes, and the weights are used for distance adjustment.

Property 2: *The interpretable directions are almost orthogonal with each other [Härkönen* et al.*, 2020].*

This property guarantees no need to train representations for the combinations of different attributes. Instead, we can train representations for attributes individually and combine them by addition. This is because adjusting the age attribute through age representation addition has less impact on gender representation (less movement on the gender direction) and vice versa. It increases the extensibility of our design to involve more kinds of attributes. Since if there are N kinds of attributes and for each attribute, there are n choices, we need to train N^n representations for combination attributes while only needing $N \times n$ for individual attributes.

According to the two properties, the process of adding attribute representations is designed to be:

$$\hat{y} = f_{x,\theta}(\overrightarrow{n} + \Sigma\alpha_i\overrightarrow{r_{gender_i}} + \Sigma\beta_j\overrightarrow{r_{age_j}}), \tag{3}$$

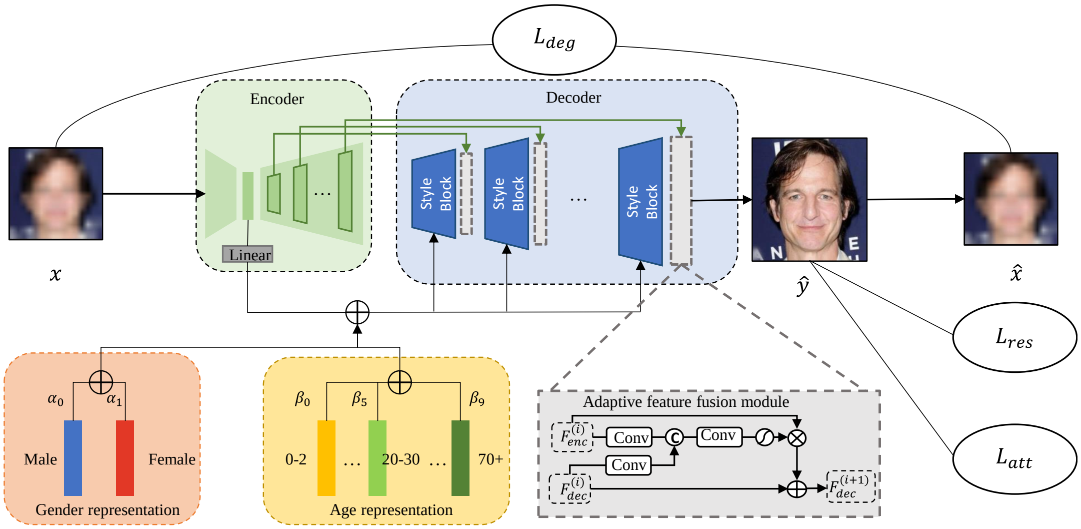


Figure 5: The architecture of DebiasFR.

where $gender_i \in \{Male, Female\}$, $age_j \in \{0-2, 3-6, 7-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-69, 70+\}$ and α_i, β_j are the weights for the representations. During training, the weights of attribute representations are decided by the attribute label of the image. The correct attribute's representation weight is set to 1, and the others are 0.

3.2 Training Strategy

Given the image and its attribute label, we train the model with two losses: *Restoration loss* and *Attribute loss*. The former requests the model to restore the image faithfully and realistically, and the latter helps form the attribute representations. A *Pseudo pair strategy* is proposed as a data augmentation strategy to help the learning of attribute representation. If the model is trained by pseudo pair, restoration loss will not be counted. Instead, a degradation loss is designed to make up for the absence of restoration loss.

Restoration Loss. The restoration loss consists of two parts: reconstruction loss and adversarial loss. We adopt pixel-wise L_1 loss and perceptual loss as reconstruction loss:

$$\begin{aligned}\mathcal{L}_{per} &= \|\phi(\hat{y}) - \phi(y)\|_1 \\ &+ \lambda_{style}\|\text{Gram}(\phi(\hat{y})) - \text{Gram}(\phi(y))\|_1, \\ \mathcal{L}_{rec} &= \lambda_{L_{pix}}\|\hat{y} - y\|_1 + \mathcal{L}_{per},\end{aligned} \tag{4}$$

where y is the ground-truth image and ϕ denotes the pre-trained VGG-19 network [Simonyan and Zisserman, 2015] and and we use the conv1, $\cdots$, conv5 feature maps before activation. The Gram$(\cdot)$ denotes calculating the Gram matrix [Gondal *et al.*, 2018]. Mapping restored the image and the target image in Gram matrix statistics can effectively reduce unpleasant artifacts [Wang *et al.*, 2021].

As for the adversarial loss, we adopt the same design as the pre-trained StyleGAN network:

$$\begin{aligned}\mathcal{L}_{adv,D} &= \mathbb{E}_{\hat{y}}[\text{Softplus}(D(\hat{y}))] + \mathbb{E}_y[\text{Softplus}(-D(y))], \\ \mathcal{L}_{adv,G} &= \mathbb{E}_{\hat{y}}[\text{Softplus}(-D(\hat{y}))],\end{aligned} \tag{5}$$

where D and G denote the discriminator and the restoration model.

Attribute Loss. The training of attribute representations is to search the interpretable direction in the latent space. To accelerate this process, we use attribute loss to constrain the attribute consistency between the input attribute weights and attribute information in the restored image. The loss can be calculated as follows:

$$\mathcal{L}_{att} = \text{CE}(a, P(a|\hat{y})), \tag{6}$$

where CE denotes the cross entropy and a is the attribute label. The probability $P(a|\hat{y})$ comes from a pre-trained classifier. The classifier's architecture consists of two parts. The first part is a pre-trained CLIP [Radford *et al.*, 2021] model, and the second part is an MLP, which classifies the extracted feature vectors. The experiments about the superiority of clip embedding are placed in the supplementary materials.

Pseudo Pair Strategy. Using the restoration loss and attribute loss, the model can only support training by the image and its correct attribute label. To alleviate the necessity of attribute labels and improve the generalized ability of the model, we propose a strategy to construct pseudo pairs for model training. The strategy helps the model train with images without ground-truth attribute labels. However, the lack of ground-truth images for calculating restoration loss can harm the model performance in image reconstruction. To address this problem, we propose *degradation loss*. The loss is designed based on the assumption that the face attribute adjustment for the restored image should be constrained by its degraded version. When the attribute of the restored image is adjusted, there should not be a big alteration for its degraded version. It needs a pair of images for model training: an HR image and its corresponding LR image. A complex degradation process synthesizes the LR image:

$$x = Deg(y) = \left((y \otimes k)_{\downarrow r} + n_\sigma\right)_{JPEG_q}. \tag{7}$$

The HR image y is blurred by convolution in the Gaussian kernel k and downsampled by a factor r. After that, additive Gaussian noise n_σ is added, and a JPEG compression with quality q is applied to obtain the LR image x. We keep the arguments of the degradation and degrade the restored image $\hat{y}$ by the same degradation procedure. Then, the degradation loss is calculated as follows:

$$\mathcal{L}_{deg} = \|Deg(\hat{y}) - Deg(y)\|_1. \tag{8}$$

The degradation process involves JPEG compression, containing some quantization operations that prevent the gradient calculation. Following the principle of straight-through estimator [Courbariaux *et al.*, 2015], we treat the quantization process as an identity function when calculating the gradient to guarantee the gradient back-propagation.

The attribute loss is also calculated on the pseudo pairs. To generate the pseudo label, we pre-train an attribute classifier predicting the labels for the LR images. The top-2 probability label will be selected. They are normalized by their sum and taken as the weight of the attribute representation and the attribute label.

4 Experimental Evaluation

4.1 Experiment Settings

Training data. We use the FFHQ-Aging dataset [Or-El *et al.*, 2020] from FFHQ [Karras *et al.*, 2019] as our training data. Compared to FFHQ, FFHQ-Aging removes images with large challenges, such as low-confidence annotation predictions, large pose variations, and severe face occlusion. It consists of 53831 images and is annotated with both gender and age. During training, all images are degraded according to Eq. (7) to obtain low-quality images. Specifically, the factor k, r, σ, q are randomly sampled from [0,0.1], [0.8,20], [0,20] and [60,100], respectively.

Test data. We use three datasets from two application scenarios, namely CelebA-HQ [Karras *et al.*, 2018], IMDB-WIKI [Rothe *et al.*, 2018], and COX [Huang *et al.*, 2015] for testing. CelebA-HQ consists of 3000 images and is widely used to evaluate the performance of face super-resolution. IMDB-WIKI and COX are two datasets that can be used for real-world face restoration applications. We clean these two datasets according to the filtering rule detailed in the supplementary materials for our experiment.

Metrics. There are two types of metrics used for evaluation. The first is the traditional image quality metrics, including PSNR, SSIM, NIQE [Mittal *et al.*, 2012] and FID [Heusel *et al.*, 2017], which measure the quality of restored images. The second is the attribute error metric, which quantifies the attribute bias in the recovered images, defined as:

$$\text{Age/Gender error} = 1 - \frac{\sum_{y \in \mathcal{D}} \mathbb{I}(\mathbf{C}(\hat{y}) \neq \mathbf{C}(y))}{|\mathcal{D}|}, \tag{9}$$

where $\mathbf{C}$ denotes the attribute classifier, which we use the pre-trained model in [Rothe *et al.*, 2018] in our experiments. $\mathbb{I}$ and $\mathcal{D}$ denote the indicator function and the image set.

Implementation details. The implementation details, like parameters setting and weights of the loss terms, are placed in the supplementary materials.

CelebA-HQ	NIQE↓	FID↓	PSNR↑	SSIM↑	Age error (%)↓	Gender error (%)↓
Bilinear	16.59	213.3	**23.25**	**0.6951**	60.90	14.37
AACNN	4.132	59.80	22.65	0.6052	43.60	2.930
PULSE	3.765	65.90	20.81	0.5695	68.90	14.16
PSFRGAN	3.982	55.88	21.71	0.6173	50.66	3.500
GFPGAN	3.800	53.87	20.72	0.6001	50.50	4.160
GPEN	3.877	**47.39**	22.12	0.6152	47.70	3.333
VQFR	**3.350**	52.09	20.48	0.5699	50.60	3.666
Ours	4.411	48.63	21.71	0.6247	**22.30**	**1.467**

Table 1: Quantitative comparison on the CelebA-HQ datasets. **Red** denotes the best performance and blue denotes the second best performance.

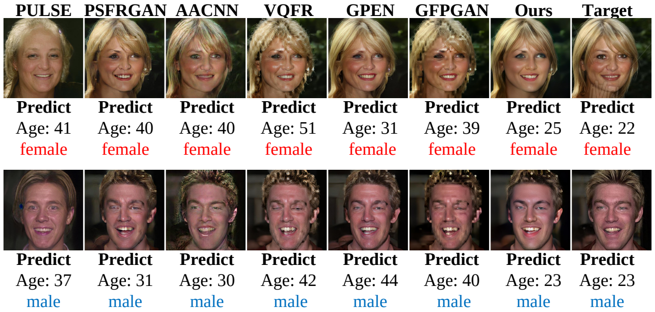


Figure 6: Qualitative comparison on the CelebA-HQ dataset. **Predict** denotes that the attribute label is obtained from model prediction. **Zoom in for best view**.

4.2 Main Results

We include five state-of-the-art methods for comparison. All experiments are performed using their official open-source models. We consider two comparison tasks: 1) 16× face super-resolution that the degradation is achieved by fixed downsampling. 2) blind face restoration that the degradation process is unknown. It is worth noting that AACNN [Lee *et al.*, 2018] is an attribute-constrained face restoration model, which also needs the attribute input but is only designed for 8× face super-resolution. We reproduce the AACNN by adding additional convolution layers with the same training data as DebiasFR to support our 16× face super-resolution task comparison.

Face Super-resolution. We conduct super-resolution experiments under a large-scale factor: 16x, which we consider a good simulation of small faces captured in surveillance scenarios. The high-quality images from CelebA-HQ and FFHQ-Aging are downsampled by bilinear interpolation and resized back to 512 × 512 to adapt to the input of the restoration model.

The quantitative results are presented in Table 1. Our method can achieve comparable face restoration results with SOTA methods and apparent alleviation in attribute bias. The qualitative results are presented in Figure 6. The images show the actual effect of our method in solving the problem of attribute bias. AACNN also receives attribute information input, but it suffers from severe artifacts. As for other methods that can generate quality images, the features like wrinkles

IMDB	NIQE↓	FID↓	Age error (%)↓	Gender error (%)↓
PSFRGAN	4.316	39.04	50.72	1.790
GFPGAN	4.133	**32.30**	48.30	1.699
GPEN	4.719	54.13	49.25	1.155
VQFR	**3.540**	33.40	50.29	1.631
Ours	4.482	39.41	**26.57**	**0.929**
COX				
PSFRGAN	4.521	88.94	61.60	21.19
GFPGAN	5.036	82.52	59.50	20.59
GPEN	4.713	84.01	64.30	22.59
VQFR	**4.190**	**70.00**	59.40	22.90
Ours	5.238	80.45	**29.50**	**8.80**

Table 2: Quantitative comparison on the IMDB-WIKI and COX datasets. **Red** denotes the best performance and blue denotes the second best performance.

and beards that do not exist in the original human face might be generated, leading to an age bias. DebiasFR circumvents this problem without harming the image quality.

Blind Face Restoration. Since the degradation model of the compared methods is slightly different from each other, instead of synthesizing LR images by our degradation model, we collect real-world LR images from the IMDB-WIKI dataset for fairness. We also compare face restoration for the images from COX, which is challenging as these images are captured under the surveillance scenario. The quantitative results are presented in Table 2. The qualitative results are presented in Figure 7 and Figure 8. It can be found that other methods may convert noise in some regions into wrinkles, whiskers, and ornaments which can cause attribute bias. Also, the repair of wrinkles can be ignored due to information loss. In contrast, DebiasFR is free from these problems.

4.3 Micro Results

Impact of Training Data Prior. Data-driven method is inevitable to be influenced by the data prior. However, we speculate that the design of attribute representations can alleviate the influence of the data prior. The training of a single attribute representation is not influenced by gradient results from other representations during the training. Also, the pseudo-pair strategy improves the generalized ability of the model for the low-frequency class images in the training set. To support our assumption, we repeat the experiments for observation 2. We propose two metrics for the experiment: *Gender var* and *Age var*. Gender var is the variance of the mean confidence of males. It evaluates the influence of the change of training set on the model. With lower variance, the model performance is more robust to the variation of training data. Age var is the variance of the age of the restored images. A large variance indicates less concentration on a single age period which implies better robustness on the influence of training data. As listed in Table 3, our method suffers from less Gender var as the gender distribution change. As for the age attribute, our method can produce a more balanced age distribution with higher variance than other methods. Also, our method suffers from lower attribute error than other methods.

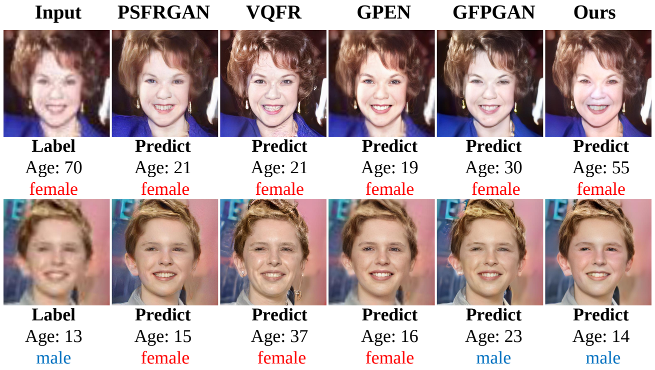


Figure 7: Qualitative comparison on the IMDB-WIKI dataset. There is no target image in IMDB-WIKI. **Label** denotes that the attribute label is offered by the dataset.

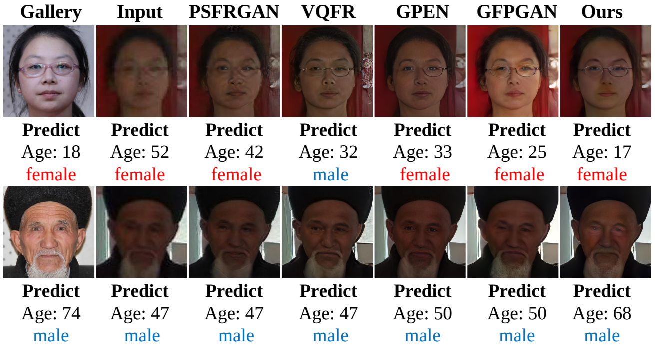


Figure 8: Qualitative comparison on the COX dataset. There is no target image but a gallery image for each person in COX.

Performance Gain of Inference Strategy. Our model accepts the weight of attributes as input, which can be obtained from either the ground-truth label or an attribute predictor estimation. We pre-trained an attribute predictor to evaluate the latter setting. Two strategies can be adopted for this two-stage inference: 1) The weight of the attribute with the largest confidence is set to 1, while the others are 0. 2) The confidences of the two most confident labels are used as weights (normalized and sum up to 1). As Table 4 shows, our model can only perform slightly better than baseline methods as the pre-trained attribute predictor performance is limited. This setting can be improved by adopting a predictor with better performance. The predictor can also predict attributes from other sources of information like voice or gait, which are extensively discussed in the context of multi-modal research.

Flexibility of Manual Control. Since the input attribute weight can be manually adjusted, the user can interact with the model by tuning the attribute weights, similar to the criminal portrait drawing. Figure 9 shows the manipulation results with different attribute weights. It can be found that the manipulation of the face mainly focuses on the high-frequency parts of the image, such as eyes, beards, wrinkles, etc. These parts are easily overlooked in image restoration and cause the attribute bias problem.

	Gender var (10^{-5})↓	Age var↑	Age error (%)↓	Gender error (%)↓
PSFRGAN	81.81	182.14	54.61	7.179
GPEN	107.38	169.47	51.92	7.948
VQFR	50.11	152.6	60.77	7.308
Ours	**8.05**	**282.6**	**34.62**	**1.667**

Table 3: Qualitative comparison on the model's robustness to the influence of data prior.

	Age error (%)↓	Gender error (%)↓
GPEN	47.70	3.33
Attribute predictor	51.43	4.13
Ours (Top1 strategy)	47.50	3.43
Ours (Top2 strategy)	46.80	3.10

Table 4: Quantitative comparison on the CelebA-HQ dataset. We only list the baseline method with the best performance in CelebA-HQ to compare with our model.

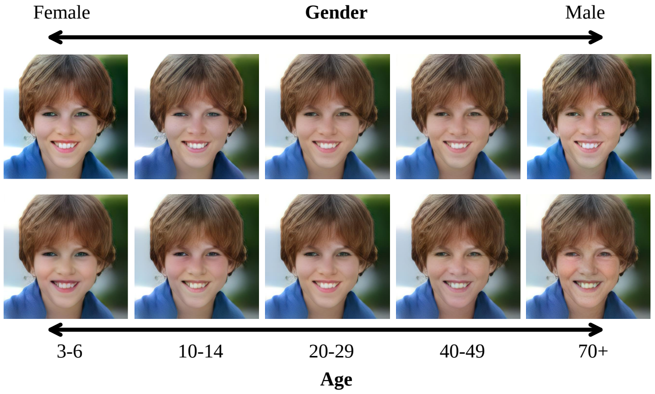


Figure 9: The manual manipulation of age and gender for an image. The first line is manipulated by setting the gender attribute weights to [2,0], [1,0], [0.5,0.5], [0,1], [0,2]. The second line is manipulated by setting the corresponding age representation weight to 1.

5 Related Work

Blind Face Restoration. Blind face restoration, also called real-world face super-resolution. Previous face super-resolution only focused on addressing the low-resolution problem, while blind face restoration should deal with the complex and unexpected degradations of the captured image in the real scene. Although the complex degradations lead to a severe loss of face information, existing works can utilize facial prior information to fill realistic face information. GAN prior and reference prior are the current mainstream used face prior. PULSE [Menon *et al.*, 2020] first proposed to leverage GAN prior to do face restoration. It redefined super-resolving a low-resolution image as generating the image whose corresponding downsampling result is most similar to it by GAN inversion. Although PULSE can produce high-quality images, it is time-consuming, and its capacity to generalize to real scenes is limited. The subsequent methods [Wang *et al.*, 2021; Yang *et al.*, 2021; Zhu *et al.*, 2022] address this problem by adopting encoder-decoder architecture and blending the features through a well-designed feature fusion module. As for the reference prior, vector quantization [van den Oord *et al.*, 2017] is the mainly used technique. The features extracted from the degraded image are used to select the suitable features from high-quality images, which help restore realistic face [Zhao *et al.*, 2022; Gu *et al.*, 2022]. Existing methods treat the face restoration task as a combination of face reconstruction and face generation. The model can generate realistic face information to replace the missing information, but the correctness of the information is not guaranteed. As a result, although existing methods have achieved great progress in restoring realistic face details, they can still suffer from the attribute bias problem. The guided face restoration methods might alleviate this problem, which require a guidance image and align the degraded image with it to achieve restoration. GFRNet [Li *et al.*, 2018] addresses this issue by training a warping sub-network. The warped image is concatenated with the degraded image and fed to the reconstruction network. ASFFNet [Li *et al.*, 2020] improves this pipeline by adopting a one-stage framework and employs spatial adaptive feature fusion. Guided-based methods should perform well when the guidance image is available and the image information perfectly aligns with the degraded image. However, it is more possible to only own simple attribute information (e.g., age and gender) in real scenes.

Attribute-constrained Face Super-resolution. Existing works have explored face super-resolution with supplementary attribute information [Yu *et al.*, 2018; Lee *et al.*, 2018; Lu *et al.*, 2018; Li *et al.*, 2019]. Although they only focus on the low-resolution problem in face restoration, these works explored the feasibility of involving attribute information. The addition of attribute information is achieved through the concatenation of image features and attribute features. These works did not yet explore the attribute bias problem, and they evaluate their methods by traditional metrics PSNR, SSIM. In our design, instead of directly fusing the image and attribute information, we map the attribute information into the latent space of GAN. Utilizing the GAN prior, our method can produce realistic face restoration. Although the editability of latent space in face attribute has been extensively studied by many face editing works [Karras *et al.*, 2019; Shen *et al.*, 2020; Härkönen *et al.*, 2020; Shen and Zhou, 2021], we are the first to propose the attribute bias problem and explore the feasibility of exploiting the editability to combat the bias problem in image restoration.

6 Conclusions

In this paper, we propose and analyze the attribute bias problem in face restoration. Existing works can restore LR images to HR images with realistic face details. However, the reconstruction of attribute information is still biased. We propose DebiasFR for combating attribute bias. The framework leverages the editability of GAN's latent space to support a fine-grained attribute manipulation on the restoration result. It is superior to SOTA solutions on attribute accuracy while with comparable image restoration performance. Furthermore, exploring the impact of DebiasFR on face recognition is an intriguing avenue for future research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China No.62206123, the Shenzhen Fundamental Research Program (Grant No.20220815112848002), and a research gift from Huawei.

References

[Chen *et al.*, 2018] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, pages 2492–2501, 2018.

[Chen *et al.*, 2021] 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*, pages 11896–11905. Computer Vision Foundation / IEEE, 2021.

[Courbariaux *et al.*, 2015] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NeurIPS*, 2015.

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

[Gondal *et al.*, 2018] Muhammad Waleed Gondal, Bernhard Schölkopf, and Michael Hirsch. The unreasonable effectiveness of texture transfer for single image super-resolution. In *ECCV*, pages 80–97. Springer, 2018.

[Grm *et al.*, 2019] Klemen Grm, Walter J Scheirer, and Vitomir Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, 29:2150–2165, 2019.

[Gu *et al.*, 2022] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, volume 13678, pages 126–143, 2022.

[Hassan *et al.*, 2021] Bilal Hassan, Ebroul Izquierdo, and Tomas Piatrik. Soft biometrics: a survey. *Multimedia Tools and Applications*, pages 1–44, 2021.

[Heusel *et al.*, 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. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NeurlPS*, pages 6626–6637, 2017.

[Huang and Liu, 2016] Dongdong Huang and Heng Liu. Face hallucination using convolutional neural network with iterative back projection. In Zhisheng You, Jie Zhou, Yunhong Wang, Zhenan Sun, Shiguang Shan, Wei-Shi Zheng, Jianjiang Feng, and Qijun Zhao, editors, *CCBR*, 2016.

[Huang *et al.*, 2015] Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen. A benchmark and comparative study of video-based face recognition on COX face database. *IEEE Trans. Image Process.*, 24(12):5967–5981, 2015.

[Härkönen *et al.*, 2020] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020.

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

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

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

[Lee *et al.*, 2018] Cheng-Han Lee, Kaipeng Zhang, Hu-Cheng Lee, Chia-Wen Cheng, and Winston Hsu. Attribute augmented convolutional neural network for face hallucination. In *CVPRW*, pages 721–729, 2018.

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

[Li *et al.*, 2019] Mengyan Li, Yuechuan Sun, Zhaoyu Zhang, Haonian Xie, and Jun Yu. Deep learning face hallucination via attributes transfer and enhancement. In *ICME*, pages 604–609. IEEE, 2019.

[Li *et al.*, 2020] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, 2020.

[Lu *et al.*, 2018] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Attribute-guided face generation using conditional cyclegan. In *ECCV*, pages 282–297, 2018.

[Menon *et al.*, 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 *CVPR*, pages 2434–2442, 2020.

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

[Or-El *et al.*, 2020] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, pages 739–755. Springer, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack

Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Rothe *et al.*, 2018] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[Shen and Zhou, 2021] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, pages 1532–1540, 2021.

[Shen *et al.*, 2020] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9243–9252, 2020.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[Song *et al.*, 2017] Yibing Song, Jiawei Zhang, Shengfeng He, Linchao Bao, and Qingxiong Yang. Learning to hallucinate face images via component generation and enhancement. In Carles Sierra, editor, *IJCAI*, pages 4537–4543. ijcai.org, 2017.

[van den Oord *et al.*, 2017] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NeurlPS*, pages 6306–6315, 2017.

[Wang *et al.*, 2021] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, pages 9168–9178, 2021.

[Yang *et al.*, 2021] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021.

[Yu *et al.*, 2018] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *CVPR*, pages 908–917, 2018.

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

[Zhao *et al.*, 2022] Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu, Changyou Chen, and Xuhui Jia. Rethinking deep face restoration. In *CVPR*, pages 7652–7661, 2022.

[Zhou *et al.*, 2015] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Learning face hallucination in the wild. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3871–3877. AAAI, 2015.

[Zhu *et al.*, 2022] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *CVPR*, pages 7662–7671, 2022.