Improving the Fairness of Deep Generative Models without Retraining

Shuhan Tan¹, Yujun Shen², Bolei Zhou²
¹Sun Yat-Sen University ²The Chinese University of Hong Kong
tanshh@mail2.sysu.edu.cn, {sy116, bzhou}@ie.cuhk.edu.hk

Abstract

Generative Adversarial Networks (GANs) have recently advanced face synthesis by learning the underlying distribution of observed data. However, biased image generation often occurs due to the imbalanced training data or the mode collapse issue. Prior work typically addresses the fairness of data generation by balancing the training data that correspond to the concerned attributes. In this work, we propose a simple yet effective method to improve the fairness of image generation for a pre-trained GAN model without retraining. We utilize the recent work of GAN interpretation to identify the directions in the latent space corresponding to the target attributes, and then manipulate a set of latent codes with balanced attribute distributions over output images. After that we learn a Gaussian Mixture Model (GMM) to fit a distribution of the latent code set, which supports the sampling of latent codes for producing images with a more fair attribute distribution. Experiments show that our method can substantially improve the fairness of image generation, outperforming potential baselines both quantitatively and qualitatively. The images generated from our method are further applied to reveal and quantify the biases in commercial face classifiers and face super-resolution model.

1. Introduction

Artificial Intelligence is being deployed in high-stakes applications in daily life, such as autonomous driving, employee hiring, loan granting, and criminal searching [13]. The decisions made by the AI algorithms, which sometimes matter to affect a person’s life, require to be unbiased and trustworthy. Unfortunately, current data-driven AI systems are known to have a discriminatory nature due to a variety of data and technical factors [2]. One major factor is the bias in the real-world training data [28]. Since current state-of-the-art AI models are mostly data-driven, the bias in the data is easily inherited by the AI models [4, 7, 5]. Such dataset bias often results in the unfair prediction of people, especially those from minority groups. Fairness, along with other ethical issues of AI algorithms, becomes a crucial issue.

Figure 1. The joint distributions of two facial attributes Age and Eyeglasses on the GAN’s training images, the images generated by the GAN, and the images generated by our proposed method respectively. We can see that the training data and the data generated by the GAN are both biased, but our method can substantially reduce such bias without retraining the model and generate more balanced images, shown below for each group. Here Age (young or adult) and Eyeglasses (with or without) are considered as binary attributes.

¹Model and demo are available at https://genforce.github.io/fairgen/.
research topic in the field.

Compared with the fairness studies on classification models [10, 23, 22, 8], relatively few works examine the fairness of the generative models. Great progress has been made by recent GAN models such as StyleGAN [14] and BigGAN [3] for synthesizing photorealistic images. As the generative models are trained with large-scale real data, they are prone to carry on the biases of the training data. As a result, many recent generative models have been found to produce biased images [34, 24, 12]. This problem becomes more concerning when these models are being used for downstream tasks. As a recent example, super-resolution method PULSE [21], which is based on a pretrained StyleGAN model [14], is found to predominantly output high-resolution white people’s faces when given low-resolution images of people from other races. This incident sparks wide concerns and discussions about the fairness of generative models [16].

Prior works have tried to improve the fairness of image generation through two approaches, the pre-processing and the in-processing: the pre-processing approach removes the bias from training data with subsampling or resampling [20], while the in-processing approach introduces new training objectives to improve data coverage and fairness [30, 25, 6, 31]. Both approaches require retraining the GAN model. However, training GAN is very costly, especially for the recent large-scale GAN models such as StyleGANs [14, 15]; meanwhile, it depends on the full access of the original training dataset, which is often not available due to privacy or storage issue. Besides, training GAN models often suffers from the mode collapse issue [27, 12], potentially bringing in more biases. To deal with the drawbacks of prior works, we explore an alternative method to improve the fairness of data generation of GAN without retraining the model.

In this work, we propose a fair image generation approach called FairGen to reduce the biases that exist in the pre-trained GANs and improve the fairness of the image generation. It does not require retraining or access to the original training data. Specifically, we first obtain a set of latent codes for synthesized images specifically for each attribute subgroup. We achieve this by shifting the distributions of sampled latent codes with a recent GAN semantic manipulation method [26], which operates in GAN’s latent space to manipulate various facial attributes. Then, we train a Gaussian Mixture Model (GMM) to fit each subgroup’s distribution in the latent space. In this way, we can sample the latent codes representing each subgroup from the GMM and then generate images with more uniformly distributed attributes using the pre-trained GAN. Fig. 1 shows that FairGen can reduce the bias of joint attribute distribution in the generated images while retaining high image qualities.

In our experiments, we show that our FairGen method can generate realistic yet diverse images for different groups, even for highly under-represented ones (occupied only 5% of the training data). More experiments on different attribute combinations show that FairGen can substantially improve the fairness of the state-of-the-art GAN models without any retraining, while outperforming the baselines both quantitatively and qualitatively. Furthermore, we use the fair image set sampled with FairGen to reveal and quantify the biases in two commercial facial classifiers and a state-of-the-art super-resolution model [21].

The contributions of this work are summarized as follows: (1) we present FairGen, a simple yet efficient method for fair attribute conditional sampling from pre-trained GAN models without retraining, (2) we analyze the bias in the generative model and show in the experiments that our method can substantially diminish the bias while preserving the image diversity, (3) we apply our method to quantify the biases in existing classification and super-resolution models.

2. Related Work

**AI fairness.** As AI systems are being used in many critical applications, AI fairness has attracted wide attention recently [2, 4, 16]. Most of the current research works focus on the fairness of the classification, where fairness is achieved when the classifier makes predictions in a non-discriminate way. Three different approaches have been proposed: 1) pre-processing methods [33, 19, 18] try to learn fair representations for the input data that benefit downstream tasks; 2) in-processing methods [29, 32, 1] add constraint or regularization during optimization of the model; 3) post-processing methods [9, 10] attempt to modify the posteriors of existing models to achieve fairness. Our method can be categorized as a post-processing approach.

**Fairness of generative models.** Zhao et al. provide one of the first studies on the bias of GAN [34], where they find that when the dataset used to train the GAN model is biased, the GAN will usually carry and amplify the bias. More recently, Jain et al. further analyzed this problem from the perspective of the mode collapse issue [12]. To reduce the bias of GAN carried from training data, several recent works have been proposed [30, 25, 6, 31, 20]. The main idea of the methods is to add additional objectives to the training process of GAN. [30, 25] aim to generate fair training datasets that could benefit the downstream classifier. They achieve this by pushing the GAN model to remove correlations between the protected attributes and the outcome label. [6] tries to offset the bias by learning a weighting function to reweigh the importance of each instance during the training. [31] mitigates bias by increasing data coverage during the training. Another way to mitigate the dataset bias is to rebalance the training
dataset directly: [20] carefully subsamples the original training set such that it is balanced w.r.t the target attributes. Although the above methods could lead to less biased GAN models, there are two main drawbacks: 1) all of them require full retraining of the GAN model, which is highly costly, especially for recent large-scale models; 2) all of these methods require the full access of the original training dataset, which is often not available for end-users. Our method overcomes these issues by focusing on the latent code sampling procedure and achieves a more fair image generation without retraining the GAN model.

3. Fair Image Generation in GANs

3.1. Problem Setting

Background. We assume there exists an unknown data distribution $P_{\text{data}}(\mathcal{X})$ observed over $d$-dimensional image data $\mathcal{X} \subseteq \mathbb{R}^d$. The goal for GAN is to learn distribution $P_\theta(\mathcal{X})$ such that $P_\theta(\mathcal{X}) = P_{\text{data}}(\mathcal{X})$. Given a well-trained GAN model, its generator can be seen as a mapping $R: \mathcal{Z} \rightarrow \mathcal{X}$ where the target fair dataset $\mathcal{D}_\text{fair}$ is sampled from $P_{\text{data}}(\mathcal{X})$, we are able to obtain a generated distribution $P_\theta(\mathcal{X}) \approx P_{\text{data}}(\mathcal{X})$. By sampling latent codes $\mathbf{z}$, we can generate a realistic dataset $\mathcal{D}_\text{direct}(g_\theta) = \{g_\theta(\mathbf{z}_i)\}_{i=1}^{N}$ with data size $N$.

Besides, we assume that each of the image sample $x$ in $\mathcal{X}$ contains specific semantic attributes, like age and gender of the face in the image. In the context of fairness, we focus on $m$ target binary attributes $\mathcal{A}_t$, which we aim to achieve fairness over their distributions. Also, we have $n'$ binary context attributes $\mathcal{A}_c$ that do not require fairness. Suppose we have a target classifier as the semantic scoring function $f_S: \mathcal{X} \rightarrow S \subseteq \mathbb{R}^m+e^m$, where $S$ is the score space for all the attributes. It can be mapped to binary labels with unit step function $H(\cdot)$, such that $H(x) = 1$ if $x \geq 0$, otherwise $H(x) = 0$. This allows us to quantify the bias in terms of specific attributes by mapping the latent code $\mathbf{z}$ to its attribute label $\mathbf{a}$ with $\mathbf{a} = H(f_S(g_\theta(\mathbf{z})))$.

Generation Bias. Real-world datasets often follow a long-tail distribution, thus carry various kinds of biases. The GAN model trained on real data $\mathcal{D}_{\text{train}}$ will therefore produce a dataset $\mathcal{D}_\text{direct}$ that highly likely carries the biases in the real data. Meanwhile, training GAN is prone to the mode collapse issue [27] which potentially brings in more biases. If the GAN models is biased, then in $\mathcal{D}_\text{direct}$ the marginal distribution of target attributes $P_{\mathcal{D}_\text{train}}(\mathcal{A}_t)$ will be highly imbalanced. In practice, we find current state-of-the-art face synthesis GAN model is indeed highly biased with $f_S$. An example is shown in Figure 1, where the rarest subgroup young people with eyeglasses is composed of only 3.0% of the generated dataset.

Measuring the Fairness of Image Generation. To reduce the bias in $\mathcal{D}_\text{direct}$, one could either acquire a balanced training set [20] or retrain the GAN model with new objectives [6, 31]. However, both approaches are costly and require access to the original training dataset and hyperparameters. Instead, we propose a new sampling strategy that could generate a fair dataset $\mathcal{D}_\text{fair}(g_\theta)$ from the same GAN model.

On one hand, our main objective is to make the marginal distribution of the target attributes in the fair dataset $P_{\mathcal{D}_\text{fair}}(\mathcal{A}_t)$ as close to the uniform distribution as possible, i.e., to reduce the divergence $D(P_{\mathcal{D}_\text{train}}(\mathcal{A}_t) \parallel U(\mathcal{A}_t))$. Here, $U$ denotes the uniform distribution. On the other hand, to ensure we do not introduce new bias w.r.t other attributes, we need to preserve the conditional distribution of the context attributes $\mathcal{A}_c$ in $\mathcal{D}_\text{direct}$. In other words, we aim to limit the divergence $D(P_{\mathcal{D}_\text{train}}(\mathcal{A}_c|\mathcal{A}_t) \parallel P_{\mathcal{D}_\text{direct}}(\mathcal{A}_c|\mathcal{A}_t))$.

Thus we formally define the fairness discrepancy $f_{gs}$ as

$$f_{gs}(\mathcal{D}_\text{fair}) = D_{KL}(P_{\mathcal{D}_\text{train}}(\mathcal{A}_t) \parallel U(\mathcal{A}_t)) + \beta D_{KL}(P_{\mathcal{D}_\text{train}}(\mathcal{A}_c|\mathcal{A}_t) \parallel P_{\mathcal{D}_\text{direct}}(\mathcal{A}_c|\mathcal{A}_t)).$$

where the target fair dataset $\mathcal{D}_\text{fair}$ is sampled from a generative model $g_\theta$ w.r.t attributes $\mathcal{A}_t$ and $\mathcal{A}_c$, $D_{KL}$ denotes the Kullback–Leibler (KL) divergence, and $\beta$ is used to balance the importance between the main objective of target fairness and the preservation constraint on context attributes. The lower $f_{gs}(\mathcal{D}_\text{fair})$ is, the more fair the generated dataset is with respect to $\mathcal{A}_t$.

3.2. FairGen Method

The goal of our method is to be able to sample a set of latent codes $\mathbf{z}_{\text{fair}} = \{\mathbf{z}_i\}_{i=1}^{N}$, such that the GAN generated dataset $\mathcal{D}_\text{fair} = \{g_\theta(\mathbf{z}_i)\}_{i=1}^{N}$ achieves low $f_{gs}(\mathcal{D}_\text{fair})$. To construct the fair latent code set $\mathbf{z}_{\text{fair}}$ that makes $P_{\mathcal{D}_\text{train}}(\mathcal{A}_t)$ close to $U(\mathcal{A}_t)$, we propose to sample a set of latent codes conditioned on each of the possible value $\mathbf{a} \in \mathcal{A}_t$.

Specifically, for $\mathcal{A}_t$ with $m$ binary attributes, we have $K = 2^m$ possible attribute values. For the $i_{th}$ possible attribute value $\mathbf{a}$, our goal is to sample a set of latent codes $\mathbf{Z}_i$ such that $H(f_S(g_\theta(\mathbf{z}_i))) = \mathbf{a}, \forall \mathbf{z} \in \mathbf{Z}_i$, and $|\mathbf{Z}_i| = N/K$. Then we can simply compose a fair latent code set with $\mathbf{Z}_{\text{fair}} = \{\mathbf{Z}_i\}_{i=1}^{K}$ such that $P_{\mathcal{D}_\text{train}}(\mathcal{A}_t) = U(\mathcal{A}_t)$. This process is illustrated in Figure 2.

The critical component of this approach is attribute-conditioned latent code sampling. We split this process into two steps: 1) Shifting Semantic Distribution: we create an intermediate code set $\mathbf{Z}_\text{edit}$ by manipulating random latent codes towards specified condition $\mathbf{a}$ with [26]; 2) Conditional Latent Space Modeling: we filter and fit the distribution of $\mathbf{Z}_\text{edit}$ with the attribute scoring function and a Gaussian Mixture Model $q_\Phi(\mathbf{z})$. Then, we can construct
Conditional Latent Space Sampling

Latent Codes
(Direct Sample)
Gaussian Mixture Model
Latent Codes
(Male)
Latent Codes
(Female)
Latent Codes
(Balanced Distribution)
Generator
Gaussian Mixture Model

Latent Space Distribution Shift
Conditional Latent Space Sampling

Figure 2. Overview of the FairGen method. For a biased GAN model, if we directly use the latent codes sampled from the initial distribution, the output images’ attribute distribution will be highly biased. We address this bias issue by first shifting the latent codes for each attribute subgroup using the identified interpretable directions corresponding to the target attributes in the latent space, and then use Gaussian Mixture Models (GMM) to fit the latent codes for each subgroup. After that, we can conditionally sample latent codes specifically from each subgroup and collect a set of latent codes with fair attribute distribution. Finally, by feeding these sampled codes to the original GAN generator, we can synthesize images with balanced attribute distribution without retraining.

3.2.1 Shifting Semantic Distribution in Latent Space

In this section, we introduce how to construct an intermediate code set \( \mathbf{Z}_{\text{edit}} \) for attribute subgroup \( \mathbf{a} \) with latent code manipulation. Specifically, we illustrate how to use InterfaceGAN [26] to map a randomly-sampled latent code \( \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d) \) to \( \mathbf{z}_{\text{edit}} \) such that \( H(f_S(g_\theta(\mathbf{z}_{\text{edit}}))) = \mathbf{a} \).

Manipulating Attributes in the Latent Space. InterfaceGAN is based on the assumption that for any binary semantic attribute \( A_{t,i} \), there exists a hyperplane in the latent space \( \mathcal{Z} \) as the separation boundary. Latent codes on the same side of the boundary have the same attribute value, which turns into the opposite when the latent codes cross the boundary.

Assume the boundary for semantic \( A_{t,i} \) has the unit normal vector \( \mathbf{n}_i \), we define its signed distance with a latent code \( \mathbf{z} \) as \( d(\mathbf{z}, \mathbf{n}_i) = \mathbf{n}_i^T \mathbf{z} \). Given \( f_i \) as the scoring function for attribute \( A_{t,i} \), when \( \mathbf{z} \) move towards or away from \( \mathbf{n}_i \), both \( d(\mathbf{z}, \mathbf{n}_i) \) and \( f_i(g_\theta(\mathbf{z})) \) would vary accordingly. Furthermore, when \( \mathbf{z} \) move across the boundary, both \( d(\mathbf{z}, \mathbf{n}_i) \) and \( f_i(g_\theta(\mathbf{z})) \) would change their numerical signs. Therefore, the authors assume a linear relationship such that

\[
\lambda d(\mathbf{z}, \mathbf{n}_i) = f_i(g_\theta(\mathbf{z})),
\]

where \( \lambda > 0 \) measures the ratio of changing speeds of semantic score and distance. According to this relationship, to manipulate the attribute score, we can easily vary the original code with \( \mathbf{z}_{\text{edit}} = \mathbf{z} + \alpha \mathbf{n}_i \), as \( f_i(g_\theta(\mathbf{z}_{\text{edit}})) = f_i(g_\theta(\mathbf{z})) + \lambda \alpha \). Please refer to Section 4.1 for the detail about how we obtain \( \mathbf{n}_i \) in practice.

With InterfaceGAN, we can edit the attribute value \( \mathbf{a} \) for sample \( \mathbf{z} \) with mapping function. Here we take the case when the \( i_{th} \) attribute \( A_{t,i} = 1 \) as an example.

Our goal is to make \( f_i(g_\theta(\mathbf{z}_{\text{edit},i})) = \lambda \alpha \), where \( \lambda \alpha > 0 \) is a predefined scoring threshold such that we can classify the synthesized image with high confidence. Towards this goal, we first move \( \mathbf{z} \) onto the decision boundary and then move it away from the boundary with magnitude \( \alpha \). Formally, we have

\[
\mathbf{z}_{\text{edit},i} = \mathbf{z} - d(\mathbf{z}, \mathbf{n}_i)\mathbf{n}_i + \alpha \mathbf{n}_i.
\]

According to (2), we can prove that \( f_i(g_\theta(\mathbf{z}_{\text{edit},i})) = \lambda \alpha \).

The next step is to set attribute value \( \mathbf{a} \) for all \( m \) target attributes \( A_{m} \). To this end, we need the normal vectors of any pair of the semantic boundary to satisfy \( \mathbf{n}_i^T \mathbf{n}_j \approx 0 \), which we find to be the case for the GAN model and attributes we consider. This is because these attributes are already well disentangled in the latent space of the pretrained model. In this way, we can set

\[
\mathbf{z}_{\text{edit}} = \mathbf{z} - \sum_{i=1}^{m} [d(\mathbf{z}, \mathbf{n}_i)\mathbf{n}_i - \alpha_i \mathbf{n}_i],
\]

where \( \alpha_i > 0 \) if \( \mathbf{a}_i = 1 \) otherwise \( \alpha_i < 0 \). It is also easy to prove that \( f_i(g_\theta(\mathbf{z}_{\text{edit}})) \approx \lambda \alpha_i \) for arbitrary attribute \( i \). In this way, if (2) is satisfied, we are able to obtain

\[
H(f_S(g_\theta(\mathbf{z}_{\text{edit}}))) = \mathbf{a}.
\]

For a specified attribute subgroup \( \mathbf{a} \), we can construct an intermediate code set \( \mathbf{Z}_{\text{edit}} \) with \( N_{\text{edit}} \) samples. This is done by simply sampling \( N_{\text{edit}} \) latent codes \( \mathbf{z} \) from \( \mathcal{N}(0, \mathbf{I}_d) \) and mapping them into \( \mathbf{z}_{\text{edit}} \) with (4).

\( \mathbf{Z} \) by sampling \( \mathbf{z} \sim q_\theta(\mathbf{z}) \). We introduce each step in more detail as follows.
3.2.2 Conditional Latent Space Modeling

To support re-sampling latent code from a specific subgroup, we utilize a Gaussian Mixture Model (GMM) to fit its distribution in the latent space with the set of latent code $Z_{edit}$ created with (4).

However, as the linear relationship in (2) may not be the case for some latent codes and attributes in practice, not all samples from $Z_{edit}$ satisfy (5). To obtain more accurate distribution modeling, we further use $f_S$ to filter out less confident samples for creating a new set: $Z'_{edit} = \{ z \in Z_{edit} | H(f_S(g_{th}(z))) = a \}$. We then train a GMM model on $Z'_{edit}$ with expectation-maximization (EM) algorithm to obtain a probabilistic model of latent codes $g_{th}(z)$ conditioned on the specified subgroup $a$. This model enables us to sample an arbitrary number of high-quality images from a certain subgroup.

To construct the fair latent code set $Z_{fair}$, we first prepare the GMM model for each of the possible value in $A_t$. We then compose $Z_{fair}$ with the same-size latent code set sampled from each of the GMM models.

4. Experiments

We evaluate FairGen with a well-trained face synthesis GAN model, StyleGAN2 [15]. Specifically, in Sec.4.1 we investigate how effective FairGen can improve the fairness of an existing GAN model w.r.t different attribute setting. Then, in Sec.4.2, to show the potential impact of FairGen, we utilize our generated fair dataset to reveal and quantify the biases in two commercial facial classification APIs as well as a state-of-the-art super-resolution model. Finally, in Sec.4.3, we conduct an ablation study of the design choices used in our model.

Implementation Details. For the GAN model $g_0$ we study, we use the official StyleGAN2 model trained on the FFHQ dataset [14]. This model is able to generate realistic human face images with $1024 \times 1024$ resolution. For the attribute classifier, we use an off-the-shelf facial classifier trained on CelebA dataset [17] as the scoring function $f_S$. Indeed this classifier itself might be biased or give some wrong predictions; for generality, we assume it is reasonably good to use.

The style-based generator in StyleGAN2 learns to first map the latent code from $Z$ space to another high-dimensional space $W$. As shown in [26], latent codes in $W$ space have much stronger disentanglement property than in $Z$, and latent code manipulation quality is also better in $W$ space. Therefore, in our framework, we directly operate and sample codes from the $W$ space, which is simply mapped from $Z$ space with StyleGAN2 mapping module.

To obtain the semantic boundaries $n$ used in (3), we first synthesize a dataset $D_{direct}$ with $50K$ images by directly sampling from the original latent space. Then, we use the scoring function $f_S$ to obtain the attribute scores for all the sensitive attributes. For each attribute, we sort the corresponding scores and choose the ones with top $2\%$ highest scores as positive examples and top $2\%$ lowest scores as negative examples. This is to select the most representative examples as the scoring function may not be absolutely accurate. Finally, we randomly select these examples into the training set to train a linear SVM to obtain the decision boundary, the normal direction of which results in $n$. The SVM is trained to take the latent codes in $W$ space as input, and output binary labels obtained with $f_S$.

For InterfaceGAN, we set the magnitude of code editing factor $|\alpha| = 3.0$. For the conditional latent space modeling of each attribute subgroup, we manipulate and generate $Z_{edit}$ with $N_{edit} = 2.5K$ samples; then we use a GMM model with $k = 10$ components to fit the distribution. We provide an empirical analysis of these hyper-parameters in Sec.4.3.

4.1. Fair Image Generation

We firstly show the existing bias in the GAN model w.r.t different attribute settings. Then, we present both the quantitative and qualitative results that show our FairGen method can significantly improve the fairness of image generation while preserving the image quality.

Experiment Setting. For evaluation, we first select 5 common facial attributes from the CelebA dataset: age, gender, smiling, eyeglasses and black_hair. To form different sampling tasks, we combine 2 of the attributes as a pair of target attributes $A_t$ to form 4 subgroups for each task while leaving the other 3 attributes in the context set $A_c$. For each of the compared sampling methods, we sample 10K images in total and evaluate on this generated dataset.

We use two metrics to evaluate the output dataset: 1) Fairness: we compute the generative sampling fairness discrepancy $f_{gs}$ with (1). Here we set $\beta = 0.1$; 2) mFID: we use the 2048-dimensional Fréchet Inception Distance (FID) [11] to evaluate the image quality and diversity. Specifically, we use $f_S$ to classify both a generated and a real dataset (FFHQ) into different sensitive attribute subgroups. Then, we compute FID only for examples from the same subgroup and obtain the average value over all subgroups.

Quantitative Evaluation. We compare FairGen with two baselines. 1) Direct Sampling: we directly sample latent codes from the original distribution of the GAN; 2) Ours w/o Edit: we classify all the images in the directly sampled dataset and directly fit the latent code of images from each subgroup with a GMM. This method reveals the effectiveness of latent code manipulation $Z_{edit}$ in Sec.3.2.1.

The results in Table 1 show that: 1) FairGen can significantly improve the fairness of GAN model in various attribute combinations. For example, for gender-eyeglasses
Figure 3. Qualitative results for fair image generation in GANs with Gender and Black Hair.

(a) Male with Black Hair
(b) Male without Black Hair
(c) Female with Black Hair
(d) Female without Black Hair

Figure 4. Qualitative results for fair image generation in GANs with Age and Eyeglasses.

(a) Young with Eyeglasses
(b) Young without Eyeglasses
(c) Adult with Eyeglasses
(d) Adult without Eyeglasses

Figure 5. Mis-classified images by APIs.

(a) Mis-classified Male images
(b) Mis-classified Female images
Table 1. Quantitative results for fair image generation in GANs

| Sensitive Attributes | age-gender | age-eyeglasses | gender-eyeglasses | gender-black_hair | age-smiling |
|----------------------|------------|----------------|-------------------|------------------|------------|
| Metrics              | mFID | Fairness | mFID | Fairness | mFID | Fairness | mFID | Fairness | mFID | Fairness |
| Direct Sampling      | 19.95 | 0.1205 | 25.71 | 0.4510 | 23.24 | 0.2808 | 29.09 | 0.2227 | 23.07 | 0.0989 |
| Ours w/o Edit        | 26.28 | 0.0298 | 30.24 | 0.0571 | 27.08 | 0.0110 | 35.05 | 0.0148 | 28.74 | 0.0182 |
| Ours                 | 27.71 | 0.0107 | 30.85 | 0.0108 | 28.02 | 0.0036 | 40.47 | 0.0057 | 29.97 | 0.0097 |

Table 2. Gender classification error rate (in percentage) in different attribute subgroups combined with gender.

| Sensitive Attributes | gender | age | eyeglasses | black_hair | smiling |
|----------------------|--------|-----|------------|------------|--------|
| Subgroups            |        |     |            |            |        |
|                      | Male   | Female | Young | Adult | With | Without | With | Without | With | Without |
| Face++ Detect API   | 0.81   | 4.19 | 2.02 | 1.76 | 0.58 | 3.40 | 0.64 | 3.64 | 3.86 | 4.06 |
| Azure Facial Recognition | 0.71 | 2.52 | 1.08 | 1.00 | 0.47 | 2.36 | 0.86 | 1.46 | 2.44 | 3.24 |

Figure 6. Examples of attribute alteration by the super-resolution model. For each set of images, from left to right we showcase 1) the original image generated by our method; 2) the LR image down-sampled from the original image; 3) HR image output by PULSE given the LR image.

Table 3. Percentage of attribute alternation by the super-resolution model.

|                | gender | glasses | age |
|----------------|--------|---------|-----|
|                | Male   | Female |     |
| gender         | 3.4    | 7.8     |     |
| glasses        | 47.0   | 45.1    | 10.8| 8.4|     |
| age            | 10.1   | 25.9    | 13.8| 15.9|     |
|                |        | w | w/o | Young | Adult |

4.2. Identifying Bias in Existing Models

The fair image generation achieved by FairGen can be useful in many applications. Here we apply our method to reveal and quantify the potential biases in the existing face classifiers and a super-resolution model.

Bias in Face Classifiers. We first study the bias in existing face classifiers. To make it more practical, we select two state-of-the-art commercial facial classification APIs (Face++ Detect API and Azure Facial Recognition). To analyze the bias problem, we focus on gender classification under different attribute conditions: age, eyeglasses, black_hair and smiling.

Specifically, for each attribute pair, we use our method to generate a fair dataset of 10K images which are evenly sampled from 4 possible subgroups (e.g., Young Male, Young Female, Adult Male, Adult Female). Then, we run the face classifiers on these datasets and compare the error rate of different groups.

We show the results in Table 2, which well quantifies the bias problem in the two APIs. In this table, besides the error rate of gender in each subgroup, we also compute...
the average error rates for male and female people over all the subgroups. We first observe that for both APIs, the gender classification accuracy for females is significantly lower than for males. Also, people without eyeglasses or black hair are more likely to be wrongly classified, while accuracy is more balanced w.r.t age and smiling. We show some of the failure cases observed by our model in Figure 5.

**Bias in Super-resolution Model.** In this part, we study the bias problem of super-resolution method PULSE [21]. This recent neural network model takes a low-resolution (LS) image as input and outputs a high-resolution (HS) image. It has been found that for certain minority groups, their rare attribute values in the LS input will often be changed to more common values in the HS output [16]. We aim to use the images generated with FairGen to examine which attributes under each subgroup will be more likely to be altered by PULSE.

To this end, we first generate images from a certain subgroup with FairGen, and then input its down-sampled LS version to PULSE to obtain a HS output. Then, we use the scoring function \( f_S \) to obtain the attribute values of the HS images. Finally, we compare this result with the original image’s attribute value, and compute the rate of PULSE alternation for each of the attributes. Here we select three attributes **Gender**, **Eyeglasses** and **Age**, and combine any 2 of them to obtain 12 subgroups in total, where each subgroup has 400 images. We exhibit some examples of the image attribute altered by PULSE in Figure 6. We can see the PULSE wrongly alters the gender, glasses and age attributes through the super-resolution.

We show the rate of attribute alternation done by PULSE in Table 3. Each column represents a certain subgroup, while each row lists the alternation rate for each attribute. We have the following fairness analysis: 1) For the gender attribute, the alternation rate is higher for female people, indicating PULSE is prone to output male faces. Also, gender is more often altered for older people. 2) For the eyeglasses attribute, we observe a very high alternate rate (89.8%) for input images with eyeglasses. This indicates that PULSE frequently fails to preserve the eyeglasses. 3) For the age attribute, it is obvious that the alternation rate for the adult group (29.5%) is much higher than in the young group (3.3%). We can conclude that PULSE is prone to output young faces. We also observe that female people’s age is more often alternated.

It is worth noting that none of the above bias analysis requires additional human labeling or careful dataset balancing for each of the attributes, making the bias analysis based on FairGen much more convenient to run than the previous bias analysis works [4, 20]. With the support to sample images from any attribute subgroup, FairGen enables us to do such detailed analysis with low cost on the biases in existing models w.r.t different subgroups. In addition, we are able to easily extend the analysis to other attributes when provided with scoring function \( f_S \) of new attributes. This makes FairGen a general and flexible tool for studying the bias of visual models.

### 4.3. Ablation Study

In this part, we analyze the impact of different design choices in FairGen on the fairness score. In particular, we analyze 1) the magnitude of InterfaceGAN manipulation \( |\alpha| \); 2) the size of edited latent code set \( N_{edit} \); 3) the number of components used in GMM models \( k \).

We plot the results in Table 7. We observe that we need a large enough \( |\alpha| \) to make sure the semantics of the shifted latent codes are correct. Secondly, the size of edited codes \( N_{edit} \) has a relatively small impact on the fairness score compared to other parameters. Finally, The more components we have in GMM, the better result we will normally obtain as more components provide a more accurate approximation of code distribution.

### 5. Conclusion

In this work, we improve the fairness of image generation in well-trained GAN models through a simple yet effective sampling-based method called FairGen. We analyze the bias in the state-of-the-art GAN models for face synthesis, and the experiments show that our method substantially reduces the bias of the facial attributes occurring in the
generated images. Furthermore, we apply the fair dataset generated by our method to reveal and quantify the biases in the face classification and super-resolution models.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Proceedings of Machine Learning Research, 2018. 2
[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. NIPS Tutorial, 2017. 1, 2
[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096, 2018. 2
[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency, 2018. 1, 2, 8
[5] Irene Chen, Fredrik D Johansson, and David Song. Why is my classifier discriminatory? In Adv. Neural Inform. Process. Syst., 2018. 1
[6] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In ICML, 2019. 2, 3
[7] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023, 2018. 1
[8] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Proceedings of Machine Learning Research, 2019. 2
[9] Michael Feldman. Computational fairness: Preventing machine-learned discrimination. PhD thesis, 2015. 2
[10] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Adv. Neural Inform. Process. Syst., 2016. 2
[11] 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 Adv. Neural Inform. Process. Syst., 2017. 5
[12] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampti. Imperfect imagination: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. arXiv preprint arXiv:2001.09528, 2020. 2
[13] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In Adv. Neural Inform. Process. Syst., 2016. 1
[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In IEEE Conf. Comput. Vis. Pattern Recog., 2019. 2, 5, 11
[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In IEEE Conf. Comput. Vis. Pattern Recog., 2020. 2, 5, 11
[16] Andrey Kurenkov. Lessons from the pulse model and discussion. https://thegradient.pub/pulse-lessons/, 2020. 2, 8
[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Int. Conf. Comput. Vis., 2015. 5
[18] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In Int. Conf. Learn. Represent., 2016. 2
[19] Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077, 2016. 2
[20] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. In Adv. Neural Inform. Process. Syst., 2019. 2, 3, 8
[21] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In IEEE Conf. Comput. Vis. Pattern Recog., 2020. 2, 8, 17
[22] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. arXiv preprint arXiv:1712.00193, 2017. 2
[23] Hee Jung Ryu, Margaret Mitchell, and Hartwig Adam. Improving smiling detection with race and gender diversity. arXiv preprint arXiv:1712.00193, 2017. 2
[24] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J Jansen. Analyzing demographic bias in artificially generated facial pictures. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 2020. 2
[25] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chentharamakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. IBM Journal of Research and Development, 2019. 2
[26] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Trans. Pattern Anal. Mach. Intell., 2020. 2, 3, 4, 5
[27] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In Adv. Neural Inform. Process. Syst., 2017. 2
[28] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In IEEE Conf. Comput. Vis. Pattern Recog., 2011. 1
[29] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Proceedings of Machine Learning Research, 2017. 2
[30] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In IEEE Int. Conf. Big Data, 2018. 2
[31] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In Eur. Conf. Comput. Vis., 2020. 2, 3
[32] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodríguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics, 2017.

[33] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In ICML, 2013.

[34] Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. In Adv. Neural Inform. Process. Syst., 2018.
Appendix

A. Data Distribution Comparisons

In this section, we show more comparisons of the data distributions of different datasets w.r.t the target attributes. Specifically, we compare the data distribution of the real training dataset FFHQ [14], the dataset directly sampled with GAN [15], and the dataset sampled with our method.

In Figure 8 we show results for the other four tasks in addition to the task we show in Figure 1 of the paper. We observe that our method is able to consistently remove the bias in the GAN model without retraining across multiple tasks.

![Data Distribution of Age - Gender](image1)

![Data Distribution of Gender - Eyeglasses](image2)

![Data Distribution of Gender - Black Hair](image3)

![Data Distribution of Age - Smiling](image4)

Figure 8. Comparisons of data distributions in different datasets.

B. Commercial API details

We detail how we test the commercial APIs in Section 4.2.

For the evaluation of facial gender classification, we first obtain access to two commercial APIs: MEGVII’s Face++ Detect API (https://www.faceplusplus.com/face-detection/) and Microsoft’s Azure Facial Recognition service (https://azure.microsoft.com/en-us/services/cognitive-services/face/). Then, we utilize FairGen to generate facial images in different subgroups. We use the gender attribute value of each subgroup as the ground-truth gender label for the images in that subgroup. After that, we use both APIs to detect and analyze the face in each image, and then compare the predicted gender attribute with the ground-truth label to obtain the accuracies in Table 2. Note that the APIs might fail to detect the face in a few images (around 0.5%), which are ignored during gender classification accuracy computation.

It is worth noting that here we are not claiming the defects or bugs in the commercial products. We just would like to raise the awareness of the potential bias in the existing applications through this small and humble academic work.
C. Conditional Generated Samples

In this section, we show more examples of the attribute subgroup images generated by our method for each task.

![Qualitative results for fair image generation in GANs with Age and Gender.](image)

(a) Young Male  
(b) Young Female  
(c) Adult Male  
(d) Adult Female

Figure 9. Qualitative results for fair image generation in GANs with Age and Gender.
Figure 10. Qualitative results for fair image generation in GANs with Age and Eyeglasses.
Figure 11. Qualitative results for fair image generation in GANs with Gender and Eyeglasses.
Figure 12. Qualitative results for fair image generation in GANs with Gender and Black Hair.
Figure 13. Qualitative results for fair image generation in GANs with Age and Smiling.
D. Bias in Super-Resolution Models

In this section, we show more examples of the failure cases we expose from the super-resolution model PULSE [21]. For each set of images, from left to right we showcase 1) the original image generated by our method; 2) the LR image subsampled from the original image; 3) HR image output by PULSE given the LR image.

Figure 14. Examples of Gender attribute alteration by the super-resolution model.

Figure 15. Examples of Eyeglasses attribute alteration by the super-resolution model.

Figure 16. Examples of Age attribute alteration by the super-resolution model.
E. Bias in Gender Classification Models

In this section, we show more examples of the failure cases of gender classification for the two APIs we show in the paper. Specifically, we show the failure cases where our generated male faces are classified as female and vice versa.

Figure 17. Mis-classified Male Images by the Commercial Gender Classification APIs
Figure 18. Mis-classified Female Images by the Commercial Gender Classification APIs