FairStyle: Debiasing StyleGAN2 with Style Channel Manipulations

Cemre Karakas*   Alara Dirik*   Eylul Yalcinkaya   Pinar Yanardag

Boğaziçi University

Istanbul, Turkey

{cemre.karakas, alara.dirik, eylul.yalcinkaya}@boun.edu.tr, yanardag.pinar@gmail.com

| Black and Female | Black and Male | Non-Black and Female | Non-Black and Male |

Figure 1. Sample outputs from the StyleGAN2 model debiased using our method with respect to **Black+Gender** attributes.

Abstract

Recent advances in generative adversarial networks have shown that it is possible to generate high-resolution and hyperrealistic images. However, the images produced by GANs are only as fair and representative as the datasets on which they are trained. In this paper, we propose a method for directly modifying a pre-trained StyleGAN2 model that can be used to generate a balanced set of images with respect to one (e.g., eyeglasses) or more attributes (e.g., gender and eyeglasses). Our method takes advantage of the style space of the StyleGAN2 model to perform disentangled control of the target attributes to be debiased. Our method does not require training additional models and directly debiases the GAN model, paving the way for its use in various downstream applications. Our experiments show that our method successfully debiases the GAN model within a few minutes without compromising the quality of the generated images. To promote fair generative models, we share the code and debiased models at http://catlab-team.github.io/fairstyle.

1. Introduction

Generative Adversarial Networks (GANs) [8] are popular image generation models capable of synthesizing high-quality images, and they have been used for a variety of visual applications [18, 28, 33, 34, 41, 42]. Like any other deep learning model, GANs are essentially statistical models trained to learn a data distribution and generate realistic data that is indistinguishable to the discriminator from that in the training set. To achieve this, GANs exploit and favor the samples that provide the most information, and may neglect minority samples. Therefore, a well-trained GAN favors learning the majority attributes, and the samples they generate suffer from the same biases in the datasets on which they are trained. For example, a GAN, trained on a face dataset with few images of non-Caucasian individuals, will generate images of mostly Caucasian individuals [21, 29]. Our preliminary analysis of the pre-trained StyleGAN2-FFHQ model confirms the significance of the generation bias: out of 10K randomly generated images, the *male* attribute is present in 42%, the *young* attribute is present in 70%, and the *eyeglasses* attribute is present in 20%. Our analysis shows that these biases also exist in the FFHQ training data with 42%, 72%, and 22% for the *male,*

---

young and *eyeglasses* attributes, respectively (see Appendix A for more details). These examples show that GANs not only inherit biases from the training data, but also carry over to the applications built on top of them. This is a particularly important issue because pre-trained large-scale GANs such as StyleGAN2 [15] are often used as the backbone of various computer vision applications in a variety of domains such as image processing, image generation and manipulation, anomaly detection, dataset generation and augmentation. Therefore, any model or application that depends on large pre-trained models such as StyleGAN2 would inherit or even amplify their biases and is therefore bound to be unfair.

In this work, we aim to address the problem of fairness in GANs by debiasing a pre-trained StyleGAN2 model with respect to single or multiple attributes. After debiasing, the edited StyleGAN2 models allow the user to generate unbiased images in which the target attributes are fairly represented. Unlike previous work that requires extensive preprocessing or training an additional model for each target attribute, our approach directly debiases the GAN model to produce more balanced outputs, and it can also be used for various downstream applications. Moreover, our approach does not require any sub-sampling of the input or output data, and is able to debias the GAN model within minutes without comprimising the image quality. Our main contributions are as follows:

- We first propose a simple method that debiases the GAN model with respect to a single attribute, such as *gender* or *eyeglasses*.

- We then extend our method for jointly debiasing multiple attributes such as *gender and eyeglasses*.

- To handle more complex attributes such as *race*, we propose a third method based on CLIP [24], where we debias StyleGAN2 with text-based prompts such as *'a black person'* or *'an asian person'*.

- We perform extensive comparisons between our proposed method and other approaches to enforce fairness for a variety of attributes. We empirically show that our method is very effective in de-biasing the GAN model to produce balanced datasets without compromising the quality of the generated images.

- To promote fair generative models and encourage further research on this topic, we provide our source code and debiased StyleGAN2 models for various attributes at http://catlab-team.github.io/fairstyle.

2. Related Work

In this section, we first review related work in fairness and bias. We then discuss studies that specifically address fairness and bias in generative models. Finally, we discuss related work in the area of latent space manipulation.

2.1. Fairness and Bias in AI

Fairness and bias detection in deep neural networks have attracted much attention in recent years [5, 22]. Most existing work on fairness focuses on studying the fairness of classifiers, as the predictions of these models can be directly used for discriminatory purposes or associate unjustified stereotypes with a particular class. Approaches to eliminating model bias can be divided into three main categories: Preprocessing methods that aim to collect balanced training data [19, 20, 40], methods to introduce constraints or regularizers into the training process [2, 36, 39], and post-processing methods that modify the posteriors of the trained models to debias them [6, 11]. In our work, we focus on debiasing and fairness methods developed specifically for GANs, which we discuss below.

2.2. Detecting and Eliminating Biases in GANs

The fairness of generative models is much less studied compared to the fairness of discriminative models. Most research on the bias and fairness of GANs aims to either eliminate the negative effects of using imbalanced data on generation results or to identify and explain the biases. Research on bias and fairness of GANs can be divided into three main categories: improving the training and generation performance of GANs using biased datasets, identifying and explaining biases, and debiasing pre-trained GANs.

The first research category, training GANs on biased datasets, aims to solve the problem of low quality image generation when the model is trained on imbalanced datasets with disjoint manifolds and fails to learn the true data distribution. [31] proposes a heuristic motivated by rejection sampling to inject *disconnectedness* into GAN training to improve learning on disconnected manifolds. [30] proposes Discriminator Optimal Transport (DOT), a gradient ascent method driven by a Wasserstein discriminator to improve samples. [3] uses a rejection sampling method to approximately correct errors in the distribution of the GAN generator. [9] proposes a weakly supervised method to detect bias in existing datasets and assigns importance weights to samples during training. The second category of research aims to detect or explain bias in generative models. [17] proposes to use attribute-specific classifiers and train a generative model to specifically explain which style channels of StyleGAN2 contribute to the underlying classifier decisions. The third line of research aims to debias and improve the sample quality of pre-trained GANs. [10] proposes to

train a probabilistic classifier to distinguish samples from two distributions and use this likelihood-free importance weighting method to correct for bias in generative models. However, this method requires training a classifier for each attribute targeted for debiasing and cannot handle biases in multiple attributes (e.g., *gender and eyeglasses*). [29] proposes a conditional latent space sampling method to generate attribute-balanced images. More specifically, latent codes from StyleGAN2 are sampled and classified. Then, a Gaussian Mixture Model (GMM) is trained for each attribute to create a set of balanced latent codes. Another recent work, [25], proposes to use the latent codes from the W-space of StyleGAN2 to train a linear SVM model for each attribute and then use the normal vector to the separation hyperplane to steer the latent code away from or towards acquiring the target attribute for debiasing. Unlike [25, 29], our method does not require model training and aims to directly debias the GAN model which can be used to generate attribute-balanced image sets.

2.3. Latent Space Manipulation

Several methods have been proposed to exploit the latent space of GANs for image manipulation, which can be divided into two broad categories: supervised and unsupervised methods. Supervised approaches typically benefit from pre-trained attribute classifiers that guide the optimization process to discover meaningful directions in the latent space, or use labeled data to train new classifiers that directly aim to learn directions of interest [7, 26]. Other work shows that it is possible to find meaningful directions in latent space in an unsupervised manner [13, 32]. GANSpace [12]) proposes to apply principal component analysis (PCA, [35]) to randomly select the latent vectors of the intermediate layers of the BigGAN and StyleGAN models. A similar approach is used in SeFA [27], where they directly optimize the intermediate weight matrix of the GAN model in closed form. LatentCLR [38] proposes a contrastive learning approach to find unsupervised directions that are transferable to different classes. In addition, both StyleCLIP [23] and StyleMC [16] use CLIP to find text-based directions within StyleGAN2 and perform both coarse and fine-grained manipulations of different attributes. Another recent work, StyleFlow [1], proposes a method for attribute-conditioned sampling and attribute-controlled editing with StyleGAN2. With respect to GAN editing, [4] proposes a method to permanently change the parameters of a GAN to produce images in which the desired attribute (e.g., clouds, thick eyebrows) is always present. However, they did not aim to debias GANs for fairness and their methodology differs from ours.

3. Methodology

In this section, we propose three methods to debias a pre-trained StyleGAN2 model. We begin with a brief description of the StyleGAN2 architecture and then describe our methods for debiasing a single attribute, joint debiasing of multiple attributes, and debiasing with text-based directions. Figure 2 illustrates a general view of our framework.

3.1. Background on StyleGAN2

The generator of StyleGAN2 contains several latent spaces: $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{W}+$ and $\mathcal{S}$, also referred to as the style space. $\mathbf{z} \in \mathcal{Z}$ is a latent vector drawn from a prior distribution $p(\mathbf{z})$, typically chosen as a Gaussian. The generator $\mathcal{G}$ acts as a mapping function $\mathcal{G} : \mathcal{Z} \rightarrow \mathcal{X}$, where $\mathcal{X}$ is the target image domain. Therefore, $\mathcal{G}$ transforms the vectors from $\mathbf{z}$ into an intermediate latent space $\mathcal{W}$ by forward propagating them through 8 fully connected layers. The resulting latent vectors $\mathbf{w} \in \mathcal{W}$ are then transformed into channel-wise style parameters, forming the *style space*, denoted $\mathcal{S}$. In our work, we use the style space $\mathcal{S}$ to perform manipulations, as it is shown [37] to be the most disentangled, complete and informative space of StyleGAN2.

The synthesis network of the generator in StyleGAN2 consists of several blocks, each block having two convolutional layers for synthesizing feature maps. Each main block has an additional 1×1 convolutional layer that maps the output feature tensor to RGB colors, referred to as *tRGB*. The three different style code vectors are referred to as $\mathbf{s}_{B1}$, $\mathbf{s}_{B2}$, and $\mathbf{s}_{B+tRGB}$, where B indicates the block number. Given a block B, the style vectors $\mathbf{s}_{B1}$ and $\mathbf{s}_{B2}$ of each block consist of style channels that control disentangled visual attributes. The style vectors of each layer are obtained from the intermediate latent vectors $\mathbf{w} \in \mathcal{W}$ of the same layer by three affine transformations, $\mathbf{w}_{B1} \rightarrow \mathbf{s}_{B1}, \mathbf{w}_{B2} \rightarrow \mathbf{s}_{B2}, \mathbf{w}_{B2} \rightarrow \mathbf{s}_{B+tRGB}$.

3.2. Measuring Generation Bias

To assess whether our method produces a balanced distribution of attributes, we begin by formulating and quantifying the bias in the generated images. Given an n-dimensional image dataset $\mathcal{I} \subseteq \mathbb{R}^n$, GANs attempt to learn such a distribution $P(\mathcal{I}) = P_{\text{data}}(\mathcal{I})$. Thus, a well-trained generator is a mapping function $\mathcal{G} : \mathcal{Z} \rightarrow \mathcal{I}$, where $\mathcal{Z} \subseteq \mathbb{R}^m$ denotes the m-dimensional latent space, usually assumed to be a Gaussian distribution. Moreover, we can sample latent codes $\mathbf{z}$ and use the trained model to generate a realistic dataset $D = \{\mathcal{G}(\mathbf{z}_i)\}_{i=1}^{N}$ of N generated images belonging to the distribution $P(\mathcal{I}) \approx P_{\text{data}}(\mathcal{I})$.

Assuming that real and generated images contain k semantic attributes $a_1, a_2, ..., a_k$, a well-trained GAN learns any bias inherent in the original data distribution $P_{\text{data}}(\mathcal{I})$ with respect to the semantic attributes. In our work, we
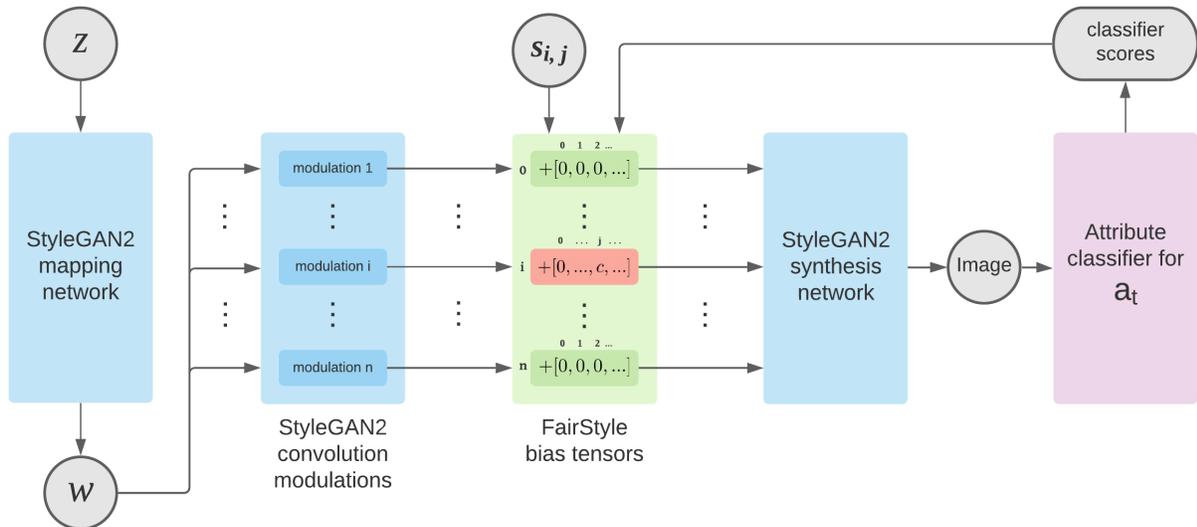
Figure 2. An overview of the FairStyle architecture, $\mathbf{z}$ denotes a random vector drawn from a Gaussian distribution, $\mathbf{w}$ denotes the latent vector generated by the mapping network of StyleGAN2. Given a target attribute a_t, $s_{i,j}$ represents the style channel with layer index i and channel index j controlling the target attribute. We introduce *fairstyle* bias tensors into the GAN model, in which we edit the corresponding style channel $s_{i,j}$ for debiasing. The edited vectors are then fed into the generator to get a new batch of images from which we obtain updated classifier results for a_t. The fairstyle bias tensors are iteratively edited until the GAN model produces a balanced distribution with respect to the target attribute. The de-biased GAN model can then be used for sampling purposes or directly used as a generative backbone model in downstream applications.

are interested in finding both the marginal distribution of the individual semantic attributes $P(a_i)$ and the joint distributions of the attribute pairs $P(a_i, a_j)$ of the generated dataset D. To measure generation bias, we generate N random images with pre-trained StyleGAN2 trained on the FFHQ dataset, and use 40 pre-trained binary attribute classifiers [14] to assign labels to each image such that $a_i = 1$ if the image contains the attribute a_i, and $a_i = 0$ otherwise.

3.3. Identifying channels that control certain attributes

For a target attribute a_t such as *eyeglasses*, we first propose a simple approach that identifies a single style channel $s_{i,j}$ responsible for controlling the target attribute, where layer and channel indices are denoted by i and j, respectively. We assume that there is a binary classifier $\mathcal{C}_{a_t}$ corresponding to the target attribute, such as pre-trained CelebA binary classifiers [14]. The identified style channel $s_{i,j}$ is then used for debiasing the GAN model with respect to single (Section 3.4) and multiple attributes (Section 3.5).

To identify $s_{i,j}$, we first generate $N = 128$ random noise vectors to obtain their style codes using StyleGAN2. Given an arbitrary style code $\mathbf{s}$, we generate two perturbed style codes by adding and subtracting a value of c at the corresponding index i and channel j. This process is repeated

for 128 randomly generated style codes, and each perturbed style code is forward propagated through the Style-GAN2 generator to synthesize images. Finally, we identify $s_{i,j}$ corresponding to the target attribute by selecting the style channel for which the perturbation causes the highest average change in classification score over the batch of $N = 128$ images:

$$\underset{i,j}{\arg\max} \frac{\sum_{k=1}^{N} |\mathcal{C}_{a_t}(\mathcal{G}(\mathbf{s} - \Delta s_{i,j})) - \mathcal{C}_{a_t}(\mathcal{G}(\mathbf{s} + \Delta s_{i,j}))|}{N}$$

(1)

where $\Delta s_{i,j}$ represents the perturbation value c, k denotes the index of the generated image, and $\mathcal{G}$ denotes the generator of StyleGAN2. In other words, we repeat the same process for each channel of the style codes and leave the values of the other style channels unchanged. In our experiments, we use the perturbation value $c = 10$.

3.4. Debiasing single attributes

Once we have identified a style channel $s_{i,j}$ that controls the target attribute a_t, we can perturb the value of the channel to increase or decrease the representation of the target attribute in the generated output. In our work, we use this intuition to edit the parameters of a pre-trained StyleGAN2

model that can be used to generate balanced outputs with respect to the target attribute a_t.

To this end, we introduce additional bias tensors, which we call *fairstyle tensors*, into the GAN model (see Figure 2). These tensors are added to the StyleGAN2 convolution modulations on a channel-wise manner. More specifically, for a fairstyle tensor, $\mathbf{b}$, we set $\mathbf{b}_{i,j} = c$ and $\mathbf{b}_{m,n} = 0$, where $m, n \neq i, j$, and c is initialized to 0. In other words, the values inside the fairstyle tensors are set to zero except for the channel indices i, j that correspond to the target attribute.

We then iteratively generate a batch of $N = 128$ latent codes and compute their updated style vectors. Given an arbitrary style vector $\mathbf{s}$, we then compute the updated vector $\mathbf{s}' = \mathbf{s} + \mathbf{b}$. We forward propagate these style vectors to generate a batch of images and compute the distribution of the target attribute using an attribute classifier. Our goal is to optimize fairstyle tensor $\mathbf{b}$ such that the images generated using the updated GAN model have a fair distribution with respect to the target attribute a_t. Similar to [29], we use the Kullback-Leibler divergence between the class distribution of a_t and a uniform distribution to compute a fairness loss value $\mathcal{L}_{\text{fair}}$, formulated as follows:

$$\mathcal{L}_{\text{fair}} = KL(P_D(a_t) \,||\, \mathcal{U}(a_t)) \tag{2}$$

where P_D denotes the class probability distributions and $\mathcal{U}$ denotes the uniform distribution. We used a one-dimensional gradient descent for optimizing fairstyle tensors $\mathbf{b}$. The updated GAN model with the optimized fairstyle tensors can then be used to generate images with a balanced distribution with respect to the target attribute.

3.5. Debiasing multiple attributes

While our first method is effective at debiasing the GAN model with respect to a single attribute such as *eyeglasses*, it does not allow for the joint debiasing of multiple attributes such as *gender and eyeglasses*. Therefore, we propose to extend our method to multiple attributes. Let a_{t_1} and a_{t_2} represent attributes that we want to jointly debias, such as *gender* and *eyeglasses*. Let s_{i_1,j_1} and s_{i_2,j_2} represent the target style channels identified by the method in Section 3.3 for attributes a_{t_1} and a_{t_2}, respectively. Similar to our first method, we iteratively generate $N = 128$ random noise vectors and their corresponding style codes. Given an arbitrary style code $\mathbf{s}$, we then compute the fairstyle tensor for the corresponding channels as follows:

$$
\begin{aligned}
\mathbf{b}_{i_1,j_1} &= x_2 \times \frac{\mathbf{s}_{i_2,j_2} - \bar{\mathbf{s}}_{i_2,j_2}}{\hat{\sigma}_{\mathbf{s}_{i_2,j_2}}} + y_2 \\
\mathbf{b}_{i_2,j_2} &= x_1 \times \frac{\mathbf{s}_{i_1,j_1} - \bar{\mathbf{s}}_{i_1,j_1}}{\hat{\sigma}_{\mathbf{s}_{i_1,j_1}}} + y_1
\end{aligned}
\tag{3}
$$

where x_1, y_1, x_2, y_2 are learned parameters initialized at 0 and optimized using gradient descent over a batch of N images, and $\bar{s}_{i,j}$, $\hat{\sigma}_{s_{i,j}}$ denote the mean and standard deviation for a given target style channel $s_{i,j}$ calculated as follows:

$$\bar{\mathbf{s}}_{i,j} = \frac{1}{N} \sum_{k=1}^{N} s_{i,j} \tag{4}$$

$$\hat{\sigma}^2_{\mathbf{s}_{i,j}} = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{s}_{i,j} - \bar{\mathbf{s}}_{i,j})^2 \tag{5}$$

Similar to our first method, we use KL divergence as a loss function between the joint class distribution of attributes a_{t_1}, a_{t_2} and a uniform distribution. After optimizing the fairstyle tensor, we use the GAN model to produce a balanced distribution of images with respect to the target attributes.

Our method can also be extended to support joint debiasing for more than two attributes. Let the number of attributes for which we want to jointly debias our model be M and assume that we have identified a style channel $s_{i,j}$ for each target attribute. In this case, each corresponding channel of the fairstyle tensor is updated as follows:

$$\mathbf{b}_{i_m,j_m} = \sum_{k=1, k \neq m}^{M} \left(x_{m_k} \times \frac{\mathbf{s}_{i_k,j_k} - \bar{\mathbf{s}}_{i_k,j_k}}{\hat{\sigma}_{\mathbf{s}_{i_k,j_k}}} + y_{m_k} \right) \tag{6}$$

We note that Eq. 6 is simply a generalized version of Eq. 3 where each fairstyle tensor channel for a target depends on the other target channels. In this case, the number of resulting subclasses is equal to M^2 and the number of parameters to be learned is equal to $2 \times M \times (M-1)$.

3.6. Debiasing attributes with text-based directions

The first two methods debias the GAN model with single or multiple channels, where the channels responsible for the desired attributes were identified using pre-trained attribute classifiers. However, the complexity of the attributes is limited by the availability of the classifiers. To debias even more complex attributes such as '*a black person*' or '*an asian person*', we debias style channels with text-based directions using CLIP. We use StyleMC [16] to identify the individual style channels for a given text.

In addition to the text-based directions, we also replace the attribute classifier with a CLIP-based one, since binary classifiers are not available for more complex attributes. In this case, we label images by comparing their CLIP-based distances D_{CLIP} with a text prompt a_t describing our target attribute and with another text prompt $a_{t_{neg}}$ negating the attribute (e.g., 'the photo of a person with curly hair' vs. 'the photo of a person with straight hair') as follows:
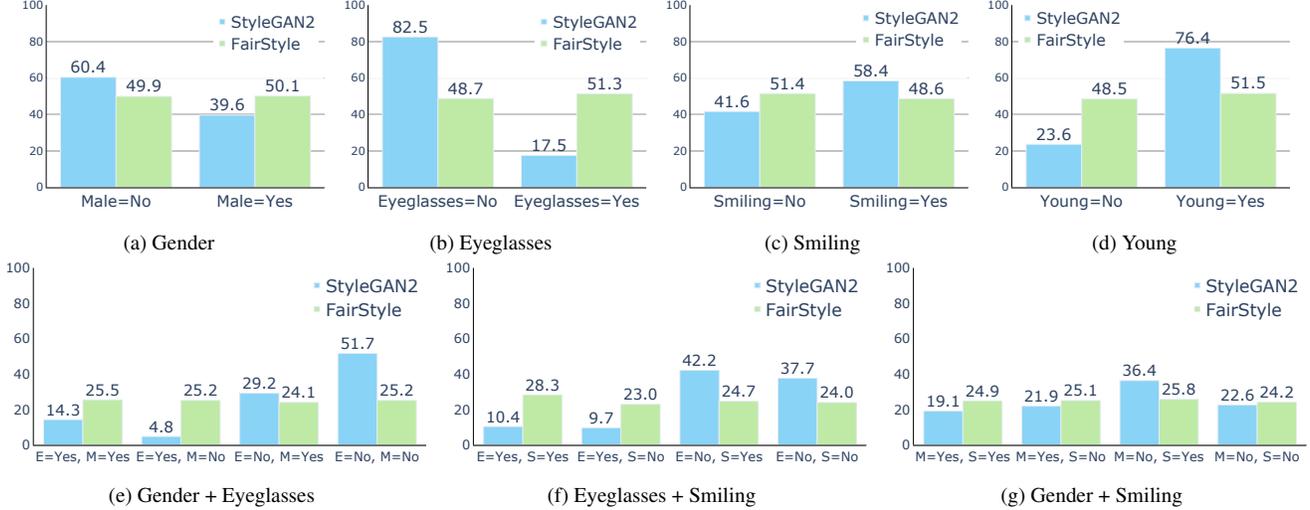
5

Figure 3. Distribution of single and joint attributes before and after debiasing StyleGAN2 model with our methods.

$$\mathcal{C}_{a_t} = \begin{cases} 1, & \text{if } D_{\text{CLIP}}(\mathcal{G}(\mathbf{s}), a_t) < D_{\text{CLIP}}(\mathcal{G}(\mathbf{s}), a_{t_{neg}}). \\ 0, & \text{otherwise.} \end{cases}$$

(7)

where $\mathbf{s}$ is an arbitrary style code, D_{CLIP} is the cosine distance between CLIP embeddings of the generated image and the text prompt a_t or $a_{t_{neg}}$, and $\mathcal{C}_{a_t}$ is the binary label assigned based on whichever text prompt (a_t or $a_{t_{neg}}$) achieves the shortest CLIP distance from the input image. We note that the negative text prompt $a_{t_{neg}}$, as in the example above, may be biased and exclude certain groups, such as *'the photo of a black person'*.

With an effective approach to assign classification scores to generated images, we identify a direction s_{a_t} consisting of one or more style channels using [16]. We use the same debiasing approach as our first method by replacing $\mathbf{b}$ with αs_{a_t}, where α is the hyperparameter for manipulation strength.

4. Experiments

In this section, we explain our experimental setup and evaluate the proposed methods using StyleGAN2 trained on the FFHQ dataset. Furthermore, we show that our methods effectively debias StyleGAN2 without requiring model training or affecting the quality of generation. Next, we compare our methods to FairGen [29] and StyleFlow [1] methods.

4.1. Experimental Setup

For the first two methods, we identify a layer and a style channel for the *gender, eyeglasses, smiling* and *age* attributes and use them in our single or multiple attribute

debiasing methods as described in Section 3.4 and Section 3.5. For the third method, described in Section 3.6, we experiment with a variety of simple and complex attributes such as *'a person with eyeglasses'*, *'a smiling person'*, *'a black person'*, *'an asian person'* using [16]. We generate and label 1000 images to compute the mean and std statistics for our second method.

For our experiments, we use the official pre-trained StyleGAN2 models and binary attribute classifiers pre-trained with the CelebA-HQ dataset[1]. To identify attribute-relevant style channels, we exclude s_{tRGB} layers from the style channel search since they cause entangled manipulations [37]. Following [16], we also exclude the style channels of the last four blocks from the search, as they represent very fine-grained features.

For the comparison with FairGen, we use the pre-trained GMM models[2]. For FairGen, we had to limit our comparison to the available pre-trained models in Table 1. We used the StyleFlow's official implementation[3] to uniformly sample latent codes from each attribute group. Although Style-Flow is not intended for fairness, we use it for conditional sampling similar to [29]. In StyleFlow, we had to limit our comparisons to *gender, smiling, eyeglasses* and *age* and their multiple attributes *age and eyeglasses*, *age and gender*, *gender and eyeglasses*. We exclude the comparison for *racial attributes* for both methods because no pre-trained models were available for these attributes or training code to train new ones.

---

[1] https://github.com/NVlabs/stylegan2
[2] https://github.com/genforce/fairgen
[3] https://github.com/RameenAbdal/StyleFlow

6

(a) Female with Eyeglasses   (b) Female w/o Eyeglasses   (c) Male with Eyeglasses   (d) Male w/o Eyeglasses

(e) Black with Eyeglasses   (f) Black w/o Eyeglasses   (g) Non-Black with Eyeglasses   (h) Non-Black w/o Eyeglasses
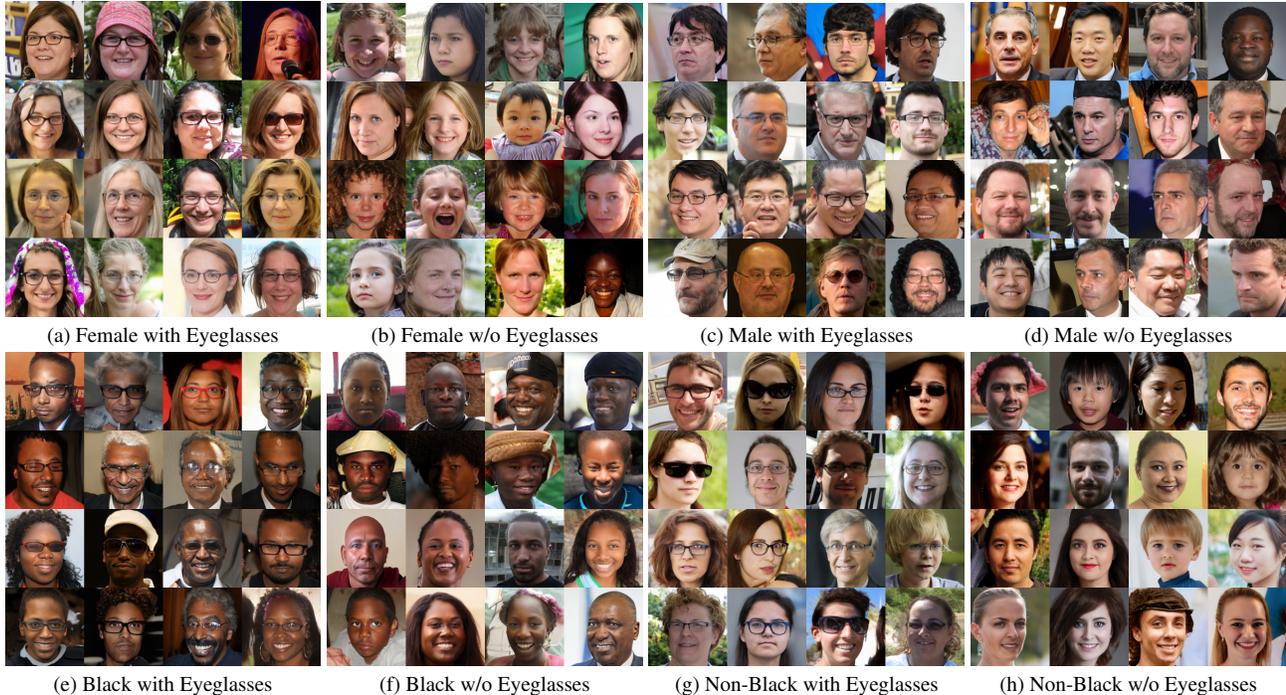
Figure 4. Qualitative results for fair image generation in GANs with **Gender+Eyeglasses** and **Black+Eyeglasses** attributes.

4.2. Fairness Analysis

To assess the fairness of the generated images, we report the KL divergence between the marginal or joint distribution of the generated images with respect to the target attributes and a uniform distribution (see Eq. 2). Our goal is to obtain a distribution with respect to one or more attributes that closely resembles a uniform distribution in order to achieve a fair distribution. To this end, we generate 10K images for each of our methods as well as for the pre-trained StyleGAN2 model, FFHQ dataset, FairGen and StyleFlow.

We start with our first method to debias a single target attribute, and present marginal distribution of the datasets generated with our method and the pre-trained StyleGAN2 in Figure 3 (a-d). As can be seen in the figure, our first method can successfully debias attributes and achieves almost perfectly balanced datasets for the attributes *gender, eyeglasses, age* and *smiling*. Next, we use our second method to debias *gender and eyeglasses, eyeglasses and smiling* and *gender and smiling* attributes. As can be seen in Figure 3 (e-g), our second method is very effective at debiasing even extremely imbalanced distributions as in the case of the *gender and eyeglasses* attributes, and can achieve a significant balance.

We then measure the KL divergence between the distribution of generated datasets and a uniform distribution, and provide a comprehensive comparative analysis

with the FFHQ training dataset, pre-trained StyleGAN2, FairGen, and StyleFlow. We debias single attributes for *eyeglasses, age, smiling, gender* and joint attributes for the Age+Gender, Age+Eyeglasses, and Gender+Eyeglasses (see Table 1). As can be seen in the table, our method outperforms StyleFlow, Fairgen and the pre-trained StyleGAN model on all attributes and achieves KL divergence values that are very close to uniform distribution in all single-attribute debiasing experiments.

We also perform additional single-attribute debiasing experiments for the highly biased attributes *black, asian,* and *white*. Since the CelebA classifiers did not cover these attributes, we used our CLIP-based method to debias the StyleGAN2 model for the *black, asian,* and *white* attributes. We present the results of this experiment in Table 2. As can be seen in the table, our method achieves a distribution that is very close to a uniform distribution, and effectively produces unbiased datasets with respect to the racial attributes.

4.3. Qualitative Results

We use our methods to debias StyleGAN2 for multiple attributes and show the generated images in Figure 1 and Figure 4. As can be seen in the figures, our method is able to generate balanced images for the attributes *gender with eyeglasses* (Figure 4 (a-d)), *gender and black* (Figure 1 (a-d)) and attributes *black and eyeglasses* (Figure 4) (e-h).

Table 1. KL Divergence between a uniform distribution and the distribution of images generated with our method, StyleFlow and FairGen. FFHQ and StyleGAN2 are included for comparison purposes.

Method	Age+Gender	Age+Glasses	Gender+Glasses	Glasses	Age	Smiling	Gender
FFHQ	0.2456	0.3546	0.2421	0.186	0.091	0.005	0.015
StyleGAN2	0.2794	0.3836	0.2495	0.180	0.109	0.011	0.018
StyleFlow	0.2141	0.1620	0.1214	0.061	3.98×10^{-4}	0.045	0.023
FairGen	3.73×10^{-2}	3.30×10^{-2}	1.85×10^{-3}	7.07×10^{-4}	1.77×10^{-3}	1.80×10^{-5}	4.21×10^{-4}
FairStyle	2.57×10^{-2}	1.57×10^{-2}	2.41×10^{-4}	0	1.80×10^{-7}	8×10^{-8}	3.20×10^{-7}

4.4. Runtime Analysis

Our method directly debias the StyleGAN2 model within a short period of time. More specifically, the average time to debias a single attribute is 2.25 minutes, while debiasing joint attributes takes 4.2 minutes.

4.5. Generation Quality

We note that a fair generative model should not compromise on generation quality to maintain its usefulness. To ensure that our methods generate high quality and diverse images, we report the Fréchet Inception Distance (FID) between sets of $10K$ images generated by the debiased Style-GAN2 model produced by our method and by the pre-trained StyleGAN2 model. Unlike our method, FairGen and StyleFlow do not edit the GAN model, but rely on subsampling latent vectors from GMM or normalizing flows models. Therefore, we exclude them from the FID experiments.

To test image quality after debiasing the GAN model, we use the attribute pairs *gender and eyeglasses*, *race and gender* and *race and eyeglasses* to compute the FID scores of the debiased datasets. While the pre-trained StyleGAN2 model achieves a FID score of 14.11, our method achieves fairly similar FID score of 14.72 (a lower FID score is better). Note that a small increase in FID scores is expected as the distribution of generated images is shifted for debiasing compared to the real images from the training data. However, we note that the increase in FID score is negligible and the debiased GAN model still generates high quality images (see Figure 1 and Figure 4).

5. Limitations and Broader Impact

While our proposed method is effective in debiasing GAN models, it requires pre-trained attribute classifiers for style code optimization. We note that the debiasing process can be affected by biases in these classifiers, a problem that also occurs in the competing methods. This is especially important when debiasing attributes that are known to be

Table 2. KL Divergence between a uniform distribution and the distribution of images generated by our text-based method to debias the *black*, *asian*, and *white* attributes. FFHQ and StyleGAN2 are included for comparison purposes.

Method	Black	Asian	White
FFHQ	0.576	0.279	0.042
StyleGAN2	0.603	0.319	0.057
FairStyle	8.00×10^{-6}	7.20×10^{-7}	2×10^{-6}

biased, such as racial attributes like *black* or *asian*.

6. Conclusion

Generative models are only as fair as the data sets on which they are trained. In this work, we attempt to address this problem and propose three novel methods for debiasing a pre-trained StyleGAN2 model to allow fairer data generation with respect to a single or multiple target attributes. Unlike previous work that requires training a separate model for each target attribute or subsampling from the latent space to generate debiased datasets, our method restricts the debiasing process to the style space of Style-GAN2 and directly edits the GAN model for fast and stable fair data generation. In our experiments, we have shown that our method is not only effective in debiasing, but also does not affect the generation quality.

We believe that our method is not only useful for generating fairer data, but also our debiased models can serve as a fairer framework for various applications built on StyleGAN2. We hope that our work will not only raise awareness of the importance of fairness in generative models, but also serve as a foundation for future research.

References

[1] Rameen Abdal, Peihao Zhu, Niloy Jyoti Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ArXiv*, abs/2008.02401, 2021. 3, 6

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. *ArXiv*, abs/1803.02453, 2018. 2

[3] Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian J. Goodfellow, and Augustus Odena. Discriminator rejection sampling. *ArXiv*, abs/1810.06758, 2019. 2

[4] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. *ArXiv*, abs/2007.15646, 2020. 3

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, 2018. 2

[6] Michael Feldman. Computational fairness: Preventing machine-learned discrimination. 2015. 2

[7] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 1

[9] Aditya Grover, Kristy Choi, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *ICML*, 2020. 2

[10] Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *DGS@ICLR*, 2019. 2

[11] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016. 2

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 3

[13] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. 3

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 4

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020. 2

[16] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: Multi-channel based fast text-guided image generation and manipulation. *ArXiv*, abs/2112.08493, 2021. 3, 5, 6

[17] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in style: Training a gan to explain a classifier in stylespace. *ArXiv*, abs/2104.13369, 2021. 2

[18] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K. Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis, 2019. 1

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 2

[20] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2016. 2

[21] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. *arXiv preprint arXiv:1906.11891*, 2019. 1

[22] L. Oneto and Silvia Chiappa. Fairness in machine learning. *ArXiv*, abs/2012.15816, 2020. 2

[23] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 3

[24] A. Radford, J. W. Kim, Chris Hallacy, Aditya Ramesh, G. Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, J. Clark, G. Krüger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ArXiv*, abs/2103.00020, 2021. 2

[25] Vikram V. Ramaswamy, Sunnis S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9297–9306, 2021. 3

[26] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[27] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 3

[28] Wanjie Sun and Zhenzhong Chen. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing*, 29:4027–4040, 2020. 1

[29] Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *ArXiv*, abs/2012.04842, 2020. 1, 3, 5, 6

[30] A. Tanaka. Discriminator optimal transport. In *NeurIPS*, 2019. 2

[31] Ugo Tanielian, Thibaut Issenhuth, Elvis Dohmatob, and Jérémie Mary. Learning disconnected manifolds: a no gans land. *ArXiv*, abs/2006.04596, 2020. 2

[32] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020. 3

[33] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson Lau. Spatial attentive single-image deraining with a high quality real rain dataset, 2019. 1

[34] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans, 2017. 1

[35] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 3

[36] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *ArXiv*, abs/1702.06081, 2017. 2

[37] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. *arXiv preprint arXiv:2011.12799*, 2020. 3, 6

[38] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. *arXiv preprint arXiv:2104.00820*, 2021. 3

[39] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017. 2

[40] Richard S. Zemel, Ledell Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013. 2

[41] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916, 2017. 1

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. 1

A. Fairness Analysis on FFHQ Data and Style-GAN2 FFHQ Model

To understand how fair the StyleGAN2 model works on FFHQ, we randomly generated 1000 images. Then we used binary classifiers to label each image for the attributes *gender, smiling, eyeglasses*, and *young* for marginal and joint distributions (Table 3, Table 4). As can be seen, the StyleGAN2 model generates images that are slightly biased towards *Male=False*, moderately biased towards *Smiling=True* and strongly biased towards *Young=True* and *Eyeglasses=False* attributes.We also examine the joint distribution of attribute pairs such as *gender + eyeglasses*, *gender + smiling* and *eyeglasses + smiling*. As can be seen, the joint probability distribution of the attributes can be extremely imbalanced even if the marginal probability distributions of the individual attributes are not, such as the ratio of *women + eyeglasses* to *men + eyeglasses*. In Figure 7 and Figure 8, respectively, we show the percentage of assigned binary labels for single and multiple attributes.

B. Additional debiasing results

We also performed debiasing for *eyeglasses* (Figure 5) and *afro hair* attribute (Figure 6) on the same latent codes showing the before/after of our debiasing method.

Figure 5. A set of images generated with the same latent codes before and after debiasing the StyleGAN2 model with respect to the 'Eyeglasses' attribute on a single channel with our method.
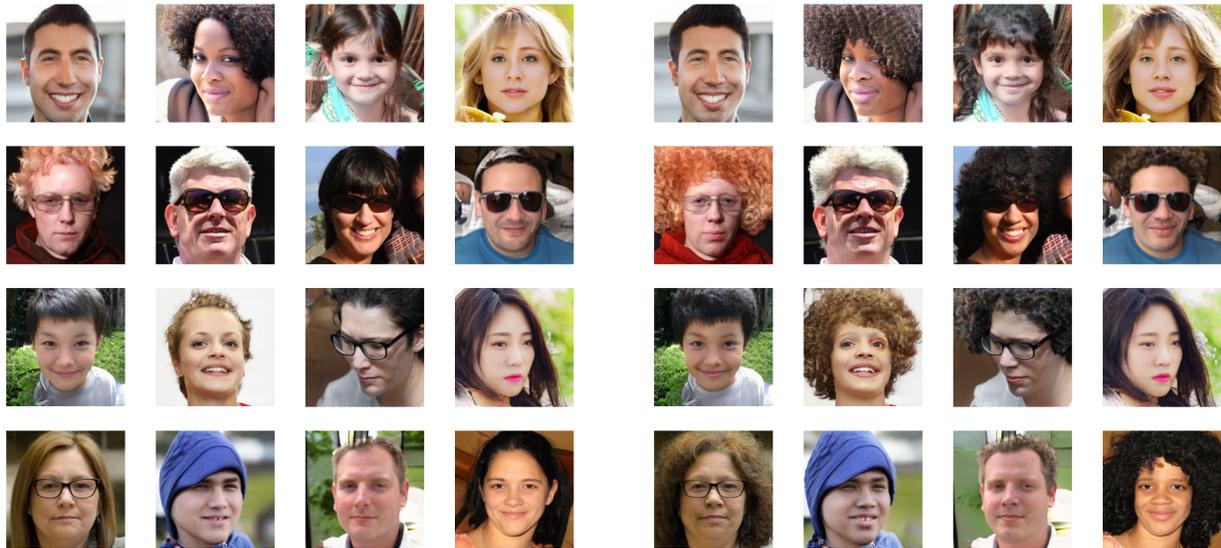

Figure 6. A set of images generated with the same latent codes before and after debiasing the StyleGAN2 model with respect to the 'a person with afro hairstyle' text-based attribute with our method.

Table 3. Marginal distributions of attributes measured on the FFHQ dataset and images generated by StyleGAN2 pretrained on the FFHQ dataset.

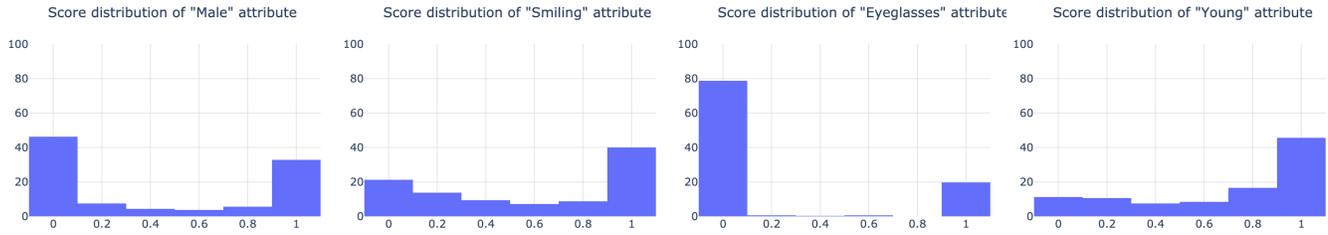| Attribute	FFHQ	StyleGAN2
Eyeglasses	F=0.78, T=0.22	F=0.80, T=0.20
Young	F=0.28, T=0.72	F=0.30, T=0.70
Smiling	F=0.43, T=0.57	F=0.44, T=0.56
Male	F=0.58, T=0.42	F=0.58, T=0.42

Figure 7. Marginal probability distributions of 'male', 'smiling', 'eyeglasses', 'young' attributes sampled from images generated by StyleGAN2 pre-trained on the FFHQ dataset.
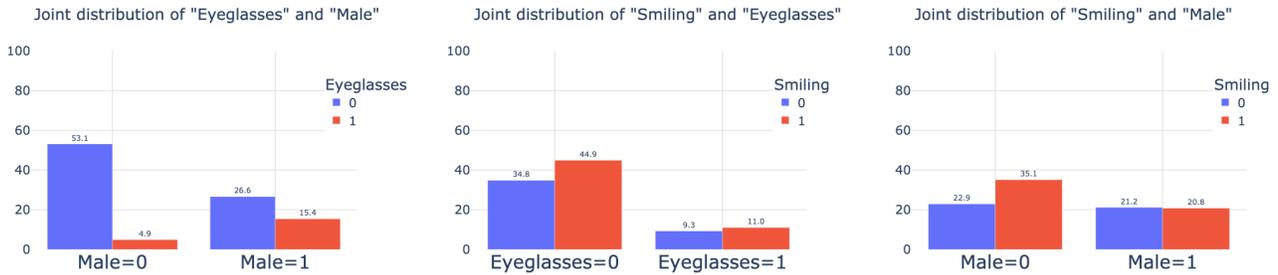

Figure 8. Joint probability distributions of ('male', 'eyeglasses'), ('eyeglasses', 'smiling'), ('male', 'smiling') attribute pairs sampled from images generated by StyleGAN2 pre-trained on the FFHQ dataset.

Table 4. Joint distributions of attribute pairs measured on the FFHQ dataset and images generated by StyleGAN2 pretrained on the FFHQ dataset.

Attributes	FFHQ	StyleGAN2
Eyegl.-Smile	FF=0.34, FT=0.44 TF=0.09, TT=0.13	FF=0.35, FT=0.45 TF=0.09, TT=0.11
Smile-Male	FF=0.22, FT=0.36 TF=0.21, TT=0.21	FF=0.23, FT=0.35 TF=0.21, TT=0.21
Male-Eyegl.	FF=0.50, FT=0.08 TF=0.28, TT=0.14	FF=0.53, FT=0.05 TF=0.27, TT=0.15