GAMMA-FACE: GAussian Mixture Models Amend Diffusion Models for Bias Mitigation in Face Images

Basudha Pal[1†], Arunkumar Kannan[1†], Ram Prabhakar Kathirvel[1], Alice J. O'Toole[2], and Rama Chellappa[1]

[1] Johns Hopkins University, Baltimore MD, USA
[2] The University of Texas at Dallas, Richardson TX, USA
{bpal5,akannan7,rprabha3,rchella4}@jhu.edu, otoole@utdallas.edu

Abstract. Significant advancements have been achieved in the domain of face generation with the adoption of diffusion models. However, diffusion models tend to amplify biases during the generative process, resulting in an uneven distribution of sensitive facial attributes such as age, gender, and race. In this paper, we introduce a novel approach to address this issue by debiasing the attributes in the images generated by diffusion models. Our approach involves disentangling facial attributes by localizing the means within the latent space of the diffusion model using Gaussian mixture models (GMM). This method, leveraging the adaptable latent structure of diffusion models, allows us to localize the subspace responsible for generating specific attributes on-the-fly without the need for retraining. We demonstrate the effectiveness of our technique across various face datasets, resulting in fairer data generation while preserving sample quality. Furthermore, we empirically illustrate its effectiveness in reducing bias in downstream classification tasks without compromising performance by augmenting the original dataset with fairly generated data.

Keywords: Diffusion Model · Bias Mitigation · Gaussian Mixture Models · Data Augmentation

1 Introduction

In deep learning (DL)-based applications of computer vision, the widespread problem of bias present difficulties across various practical domains, including biometrics [28], autonomous vehicles [47], and medical imaging [29]. Specifically, in face recognition systems, deep networks are often employed to discern subject identities, yet they manifest discriminatory tendencies in terms of performance across various sensitive attributes such as age, skintone, gender, and race [23,48]. In recent years, several research works have explored the performance of face recognition systems under various circumstances, including differences in skintone, hairstyles, and makeup usage [2]. For instance, Albeiro *et al.* demonstrated

[†]Indicates equal contribution.

the effects of gender-balanced training data on face recognition and verification tasks [1], concluding that this practice does not necessarily alleviate bias in facial verification. Additionally, Buolamwini *et al.* revealed bias in commercial gender classifiers towards specific phenotypic subgroups, particularly in terms of gender and skin tone [5]. Prior efforts have also sought to mitigate gender and skin tone bias in face recognition and verification tasks using adversarial and distillation-based methods [8, 9].

These works recognize the persistent challenges of bias in computer vision, underscoring the necessity to comprehend and rectify biases in diverse contexts. Recent strides have highlighted data augmentation as a promising approach to mitigate dataset bias [24], allowing models to effectively generalize across variations in real-world data by exposing them to a diverse range of scenarios during training. While conventional data augmentation techniques offer some benefits, they have limited diversity [41]. Generative models have transformed this domain to generate images that are diverse and produce data closely resembling real-world examples [4, 12, 21, 32, 49]. These models find applications in numerous domains, including image synthesis [36, 42] text generation [14, 37], and data augmentation [3, 6]. Prior works have utilized Generative Adversarial Networks (GANs) for bias reduction outperforming standard augmentations. Recently, diffusion models [15] have become increasingly recognized for their ability to surpass GANs in terms of image generation quality [10], and their capacity to overcome certain limitations of GANs such as mode collapse and unstable training convergence. While diffusion models excel in the realm of image generation, their utility in downstream tasks such as classification and recognition remains hindered due to inherent/amplified biases in the generated data [33].

Past studies on Generative Adversarial Networks (GANs) have examined strategies for reducing bias. These strategies involve altering the latent space associated with sensitive attributes, maintaining the desired attribute unchanged. This technique [34] depends on the latent space's semantic properties. However, this method is not directly applicable to the stochastic latent space characteristic of diffusion models. Building on this foundation, we introduce a novel yet simple technique- GAMMA-FACE for reducing bias in face generative networks, utilizing Gaussian Mixture Models (GMMs) [35] to disentangle the latent space of diffusion models. Our approach utilizes the flexible nature of diffusion models' latent spaces, where each step of the sampling process is governed by a unique Gaussian distribution. Our goal is to facilitate the creation of unique and *diverse* samples for particular downstream tasks by segmenting the latent space into distinct components. Each component corresponds to disentangled latent code, allowing for equal sampling from each to produce images that not only demonstrate a controlled correlation but also more accurately reflect the original data distribution. By incorporating the data generated through this debiasing process back into the primary training dataset, we seek to reduce bias across multiple downstream classification tasks. The key contributions of our paper are as follows:

- We introduce a novel methodology for disentangling the latent space of diffusion models using GMMs. Through extensive experiments, we determine the optimal number of mixture components in the latent space of diffusion models using quantitative metrics such as the Bayesian Information Criterion (BIC). This approach represents a significant advancement by effectively disentangling latent spaces with multiple complex attribute correlations. We hypothesize that this disentanglement captures the most differentiable attributes, such as age, gender, and race and by employing GMMs, we aim to enhance the interpretability and utility of diffusion generated data in downstream classification tasks.
- We sample evenly from the components of the GMM to obtain a "fair" synthetic dataset and utilize pre-trained classifiers to create pseudo-labels for attributes like age and gender, leveraging their confidence scores. This process allows us to obtain protected attribute labels of the generated data, enabling experiments to assess their impact on other downstream tasks as we demonstrate. Our method does not require retraining as it is employed in the evaluation phase of the diffusion model. We also have an advantage over conditional diffusion models as we need to generate images just once for all of our downstream classification tasks.
- Through augmentation of original training datasets with debiased generated images, we observe a significant reduction in bias scores across multiple attributes while simultaneously improving performance in downstream classification tasks. For instance, in the case of the FFHQ dataset, for smile classification task, we achieve a 65% decrease in bias concerning age and gender, while increasing the overall accuracy of the classifier from 93.08% to 94.84%. We show that this trend is consistent across various downstream tasks and datasets, underscoring the dual effectiveness of our approach in mitigating bias as well as enhancing performance.

2 Related Work

2.1 Generative Models for Faces

Recent advancements in face image synthesis have been reshaped by the emergence of generative models, notably with the advent of generative adversarial networks (GANs) [12]. GANs, comprising a generator and a discriminator trained simultaneously, have long been at the forefront of various applications, including synthetic face generation [11,19,25,38,51]. However, GANs encounter challenges such as training difficulty and mode collapse. In contrast, denoising diffusion probabilistic models (DDPMs) offer a promising alternative. DDPMs are represented as a Markov chain with Gaussian transition distributions, employing a forward process where data transitions to isotropic Gaussian noise over diffusion time steps and a reverse process that removes noise to generate samples aligned with the training data distribution. DDPMs deliver superior quality results with simpler objectives and stable convergence, surpassing GANs in multiple applications, including face generation [10, 16, 30, 43]. Given the rising prominence of

DDPMs in facial image generation, we leverage these models to investigate the fundamental problem of bias in this domain.

2.2 Addressing Bias in Face Generative Models

In computer vision, bias remains a persistent challenge across various applications. With the rise of face generative methods based on GANs and DDPMs, concerns about bias have surfaced. Jain *et al.* [17] revealed how GANs amplify biases in gender and skin tone, while Maluleke *et al.* [27] studied biases related to racial composition in GANs. Luccioni *et al.* [26] introduced an approach to measure biases in text-to-image systems, highlighting over-representation of certain demographics. Naik *et al.* [31] examined biases in profession depiction by text-to-image models regarding gender, age, race, and location. Recently, Perera *et al.* [33] analyzed bias in unconditional face generation with diffusion models, finding amplification of age, gender, and race biases across datasets. Various methods have been explored in GANs for mitigating inherent biases [13, 44]. Grover *et al.* [13] addressed bias in deep generative models through likelihood-free importance weighting, employing a trained probabilistic classifier to estimate the unknown likelihood ratio between model and true distributions. Choi *et al.* [7] introduced a technique targeting dataset bias in deep generative models by utilizing weak supervision from a small, unlabeled reference dataset, facilitating efficient learning and generating test data aligned with the reference dataset distribution. A simpler transfer learning-based approach was adopted by researchers in [45] to achieve similar outcomes. However, these approaches are primarily geared towards reducing bias in GANs and necessitate retraining. To the best of our knowledge, prior studies have not investigated bias mitigation in diffusion models for face generation without retraining. We propose an initial exploration of using Gaussian Mixture Models (GMMs) on the latent space of DDPMs to leverage the inherent Gaussian structure for this purpose.

3 Proposed Method

Our approach, GAMMA-FACE shown in Figure 1, addresses bias in downstream classification tasks related to sensitive attributes such as gender, race, and age. We utilize GMM in the noisy latent space of diffusion models to segregate features in a high-dimensional space and generate images by sampling uniformly from each component. These images are assigned pseudo-labels for attributes by pre-trained classifiers. Additionally, for downstream tasks like smile, glasses, or hair color classification, we augment original training data with generated data and their pseudo-labels. Our results demonstrate improved overall accuracy and reduced bias (with respect to the protected attributes) in target classification tasks.

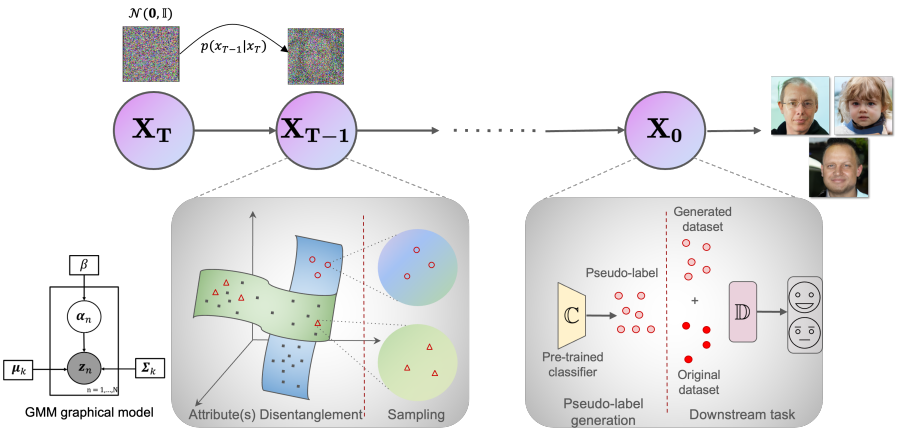

Fig. 1: The reverse diffusion phase of an unconditional diffusion model where we use a Gaussian Mixture Model (GMM) with parameters μ_k, Σ_k, and π as represented in the graphical model to disentangle latent codes into K components . Images are sampled uniformly from each component, and pre-trained classifiers ($\mathbb{C}$) provide pseudo-labels for attributes. To mitigate bias and improve performance, we retrain our downstream classifier ($\mathbb{D}$) using original image-label pairs (dark red dots) along with generated images and pseudo-labels for the target attribute (lighter red dots).

3.1 Problem Setup

Our datasets consist of facial images annotated with specific attributes, which include age (a), gender (g), and in some cases race (r). We call these protected attributes (A_p) as we aim to remove biases with respect to these attributes in our downstream tasks. Each facial image in the training set is denoted as x_i, $i \in (1 \cdots N)$, where N is the size of the dataset. The associated A_p labels for each image are denoted as a_i, g_i, and r_i. In this work, we demonstrate the effectiveness of our method on a few downstream ($\mathbb{D}$) classification tasks which include target attributes (A_t) such as smiling (s_i), presence of glasses (gl_i), hair color (h_i) classification and in some cases r_i (when we have race as the downstream task, our A_p are only age and gender).

Our approach utilizes unconditional diffusion models to generate facial images x_i through $G(\mathbf{z}_i; \theta)$, where $\mathbf{z}_i$ represents the latent code corresponding to the i^{th} image, and θ denotes the diffusion model parameters. We employ a GMM to partition the latent space $\mathbf{Z}$ into K distinct groups. This disentanglement promotes fairness by ensuring uniform sampling from each separated component, leading to the generation of new images, $\tilde{x}_i$. These generated images are then processed through a pretrained classifier ($\mathbb{C}$) to obtain pseudo-labels for A_p and A_t. We use the generated images with their A_t pseudo-labels to augment the original training dataset for the A_t classifier, aiming to enhance the classifier's performance while simultaneously reducing bias associated with A_p.

3.2 Unconditional Image Generation

Diffusion models belong to a category of generative models designed to represent the data distribution, $p_\theta(\mathbf{x}_0)$, as $\int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T}$. In the diffusion process, also known as the forward process, Gaussian noise is incrementally introduced to the data, gradually transforming the initial data point $\mathbf{x}_0$ into a state resembling pure Gaussian noise denoted as $\mathbf{x}_T$. This transformation is controlled by a sequence of variance parameters $\beta_1, \ldots, \beta_T$:

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \tag{1}$$

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{2}$$

Reversing this forward process allows the generation of new data $\tilde{\mathbf{x}}_0$ by initializing from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$. This reverse process is modeled as a Markov chain, where each step involves a learned Gaussian transition, characterized by parameters $(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$:

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \tag{3}$$

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{4}$$

Training diffusion models involves minimizing the variational bound of the negative log-likelihood of $p(x)$. The commonly used optimization objective L_{DM} reparameterizes the Gaussian transition as $\epsilon_\theta(\cdot)$ and adjusts the variational bound dynamically to improve sample quality:

$$L_{\mathrm{DM}}(\theta) := \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right\|^2 \right], \tag{5}$$

where $\mathbf{x}_t$ can be directly approximated by $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, with $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$. To sample data $\mathbf{x}_0$ from a trained diffusion model $\epsilon_\theta(\cdot)$, we iteratively denoise $\mathbf{x}_t$ from $t = T$ to $t = 1$ with noise $\mathbf{z}$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z} \tag{6}$$

3.3 GMMs for Clustering

Our primary aim is to identify distinct clusters within the latent space corresponding to various attribute values. We are not aware of the exact attributes at this point but we hypothesize that if the GMM is able to cluster the latent codes then the visually obvious attributes are segregated. To accomplish this, we employ GMMs, strategically leveraging the latent codes generated by the diffusion

model in an intermediate time step of the reverse Markov process. These latent representations during the reverse process encapsulate the high-dimensional features contributing to the synthesis of face images with specific attributes. GMM is a probabilistic model assuming data generation from a mixture of Gaussian distributions. We utilize this assumption due to the Gaussian nature of the generated latent codes, and the latent space representations of the generated face images serve as input data for the GMM.

Let $\mathbf{Z} \in \mathbb{R}^{N \times d}$ denote the latent space of the diffusion model, where N is the number of generated (latent) samples, and d is the dimensionality of the latent space vector at time step t. Each row of $\mathbf{Z}$ corresponds to a vectorized latent representation of a face image indexed by $\mathbf{z}_i$. We train a GMM to model the distribution of $\mathbf{Z}$. Formally, we model the probability distribution of latent space $\mathbf{Z}$ as a mixture of Gaussians:

$$p(\mathbf{Z}; \boldsymbol{\gamma}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{7}$$

where K is the number of components in the GMM, π_k is the weight of the k-th mixture component, and $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ is a multivariate Gaussian distribution modeling the k-th mixture component. We constrain the parameters of a GMM to be $0 \leqslant \pi_k \leqslant 1, \quad \sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\gamma} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^{K}$.

The primary objective of GMM is to determine $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ associated with each Gaussian mixture. This is achieved through the expectation-maximization (EM) algorithm. EM is an iterative method comprising two steps: the E-step computes the expectation of the log-likelihood (posterior) based on current parameters, and the M-step maximizes the GMM parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ using the posterior probability found in the E-step. Typically, the current parameters or seed values such as mean and precision matrix are initialized using a clustering framework such as k-means. We chose k-means++ as our initialization method due to its faster convergence rate. Each EM iteration increases the evidence lower bound (ELBO) to converge to the log-likelihood function of observed data. Convergence is typically assessed by monitoring the log-likelihood, stopping when a certain threshold ϵ is reached, or after a predefined number of steps. Formally, the update equations in the E-step and M-step are summarized as follows:

$$\text{E-step} : \tau_i^k = \frac{\pi_k \mathcal{N}\left(\mathbf{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)} \tag{8}$$

$$\text{M-step} : \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \tau_i^k \mathbf{z}_i}{\sum_{i=1}^{N} \tau_i^k} \tag{9}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} \tau_i^k \left(\boldsymbol{z}_i - \boldsymbol{\mu}_k\right)\left(\boldsymbol{z}_i - \boldsymbol{\mu}_k\right)^{\top}}{\sum_{i=1}^{N} \tau_i^k} \tag{10}$$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \tau_i^k \tag{11}$$

where τ_i^k represents the update equation of the E-step. In E-step, we compute the expectation of posterior distribution given the data $\mathbf{z}_i$. During M-step, we compute the desired mean vector ($\boldsymbol{\mu}_k$) and covariance matrix ($\boldsymbol{\Sigma}_k$) associated with component k using the posterior probability $\left(\tau_i^k\right)$ computed from E-step. The update process repeats until algorithm convergence, typically when the model parameters exhibit negligible changes between iterations.

A crucial hyperparameter in our GMM framework, which significantly influences model performance, is the optimal number of components K needed to effectively separate the latent space $\mathbf{Z}$. This optimal number is dependent on the characteristics of the input data. To determine the accurate count of components, we choose to employ the Bayesian Information Criterion (BIC) as our primary quantitative metric [22, 39]. We constrain our exploration of component numbers within the range of 0 to 11 and undertake a comprehensive grid search to determine the appropriate number of components. This foundational approach ensures that we identify the optimal configuration for our model, enhancing its ability to recognize and segregate the latent space into meaningful clusters corresponding to different attributes.

3.4 Application in Downstream Tasks

Diffusion models are often utilized to generate images that enhance the accuracy of various tasks. In our study, we focus on their application in classification tasks with small datasets, which tend to exhibit a noticeable bias towards specific attributes. To mitigate this bias, we sample an equal number of images from disentangled components of the GMM latent space, characterized by $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, to generate a diverse set of images.

We employ pre-trained classifiers, capable of classifying face images based on attributes A_p and A_t, to assign pseudo-labels to the generated images. These pseudo-labels with their corresponding generated images, and original image-label pairs are used to retrain a downstream A_t classifier $\mathbb{D}$. This process enhances the classifier's robustness and generalization ability across diverse inputs.

To evaluate the reduction of bias in the A_t classifier, we utilize established bias metrics from the literature. Specifically, we assess whether the classifier demonstrates any preferential bias towards a subgroup defined by an A_p. Our objective is to show that our methodology makes the classifier fairer by diminishing its bias towards certain attributes while also improving its performance.

3.5 Bias Evaluation Metrics

Our objective is to reduce the biases in downstream target classifiers aiming to enhance their overall efficacy independent of A_p. Drawing upon the foundational works in bias research [34], [8], we evaluate the effectiveness of our approach using four distinct metrics tailored to quantitatively measure the bias towards A_t in the context of A_p: Bias(A_t), the Bias Performance Coefficient (BPC), Bias Amplification (BA(A_t)), and the Kullback-Leibler (KL) divergence.

Bias: Dhar *et al.* [8] identified bias as a metric of classifier efficacy, particularly

noting the disparity in the True Positive Rate (TPR) when distinguishing images as belonging to true class or not. Sharmanska *et al.* [40] quantified bias as the weighted mean difference in false positives and false negatives across pairs of A_p. To suit our context, we modify these metrics, redefining them as the weighted average disparity in TPR across pairs of A_p for a specified A_t.

$$\text{Bias}(A_t) = \frac{\sum_{i=1}^{A_{p,\cdot}} \sum_{j=i+1}^{A_{p,\cdot}} (TPR(A_{p,i}) - TPR(A_{p,j})) \cdot w_i \cdot w_j}{\sum_{i=1}^{A_{p,\cdot}} \sum_{j=i+1}^{A_{p,\cdot}} w_i \cdot w_j} \tag{12}$$

where w_i, w_j signify the weights assigned to a specific combination of $A_{p,\cdot}$.

Bias Performance Coefficient (BPC): We further evaluate the efficacy of our technique in the context of achieving trade-off between measuring bias and classifier performance. We use Bias Performance Coefficient (BPC) proposed in [8] which measures the difference of $\text{Bias}(A_t)$ and $\text{TPR}(A_t)$ between original and de-biased datasets. High BPC signifies high bias reduction.

$$\text{BPC}(A_t) = \frac{\text{Bias}^{(A_p)} - \text{Bias}_{deb}^{(A_p)}}{\text{Bias}^{(A_p)}} - \frac{\text{TPR}^{(A_p)} - \text{TPR}_{deb}^{(A_p)}}{\text{TPR}^{(A_p)}} \tag{13}$$

Bias Amplification (BA): We use the Bias Amplification (BA) metric proposed in [46] to quantify the extent to which an A_t is associated with an A_p beyond its true occurrence rate. Let $P_{A_t|A_p}$ denote the probability of images with the A_p that also have A_t, let $P_{\hat{A}_t|A_p}$ represent the probability of images with A_p predicted to have A_t, and let P_{A_t,A_p} be the joint probability of images with both A_p and A_t. Furthermore, let P_{A_t} and P_{A_p} denote the marginal probabilities of A_t and A_p, respectively. A negative BA value suggests a reversal in bias compared to the training dataset.

$$\text{BA}(A_t) = \begin{cases} \frac{(P_{A_t|A_p} - P_{\hat{A}_t|A_p})}{P_{\hat{A}_t|A_p}} & \text{if } P_{A_t,A_p} > P_{A_t} P_{A_p} \\ -\frac{(P_{A_t|A_p} - P_{\hat{A}_t|A_p})}{P_{\hat{A}_t|A_p}} & \text{otherwise} \end{cases} \tag{14}$$

KL Divergence: This metric proposed in [34] is the Kullback-Leibler (KL) divergence to measure the disparity between score distributions across different thresholds. S_{A_p,A_t}, represents the normalized probabilities of classifier scores derived from a smoothed histogram for both A_p label and A_t label. For each A_t, the divergence is quantified by the sum of the KL divergence of the score distributions from $S_{A_p=1,A_t}$ to $S_{A_p=-1,A_t}$ and from $S_{A_p=-1,A_t}$ to $S_{A_p=1,A_t}$. This captures the divergence in classifier score distributions for samples with and without A_p, thereby offering insights into the notion of equalized odds.

$$\text{KL}(A_t) = KL[S_{A_p=-1,A_t} || S_{A_p=1,A_t}] + KL[S_{A_p=1,A_t} || S_{A_p=-1,A_t}] \tag{15}$$

4 Experiments

In this section, we detail the experimental setup that validates the efficacy of GMM-based diffusion models incorporated within the data augmentation

pipeline. This integration aims to mitigate the bias associated with A_t relative to A_p across multiple downstream classification tasks. We empirically demonstrate the capability of our approach to diminish bias by constructing an enriched training dataset that combines the original dataset with synthetically generated data. The experiments are designed to showcase the reduction in bias and the improvement in generalization performance of classification algorithms trained on this dataset against a 'Baseline' that is trained only using original data.

4.1 Datasets and Settings

We verify our experiments on two popular datasets of faces, one of high quality and the other of lower quality.

- **FFHQ Dataset:** The Flickr Faces HQ (FFHQ Dataset) is a commonly utilized dataset released by NVIDIA for tasks involving face generation [20]. This dataset comprises 70,000 aligned and cropped images with a high resolution of 1024×1024 pixels, obtained from Flickr.
- **FairFace Dataset:** The FairFace (FF) dataset, introduced in [18] consists of 108,501 aligned and cropped facial images. This dataset spans seven racial/ethnic groups, a wide range of ages from 0 to above 70 and encompasses both genders. It specifically focuses on images of non-public figures to minimize selection bias, aiming to improve fairness and accuracy in face recognition systems by providing a more diverse dataset.

In the FFHQ dataset, we group the individual ages into 'Young' for individuals under 40 and 'Old' for those 40 and older. We assign the label 'smile' to faces where the smile intensity exceeds 0.5 and 'no smile' otherwise. The presence of glasses is indicated with 'Glasses' or 'No Glasses.' Hair color is classified into 'Black,' 'Brown' (including 'Blonde'), and 'Other' (encompassing 'Green,' 'Blue,' 'Red,' 'Bald,' and other non-standard colors). Similarly, in the FF dataset, we consolidate nine age categories into 'Young' (ages 0-59) and 'Old' (ages 60 and above). We also simplify seven racial categories into 'White,' 'Black,' and 'Other' (merging 'Indian,' 'East Asian,' 'Southeast Asian,' 'Middle Eastern,' and 'Latino'). Despite balanced inputs, bias may persist, particularly in smaller datasets. Therefore, our experimental step involves creating smaller, balanced datasets from FFHQ and FF for sensitive attributes used in subsequent classification tasks. In FFHQ, we focus on classifying A_t like smile, glasses, and hair color, with age and gender as A_p. In FF, limited labels restrict our tasks to classifying one of age, race, or gender, treating the remaining as protected.

4.2 Off-the-shelf Protected Attribute Classifiers

Utilizing ResNet-101 as a standard backbone architecture, we apply pretrained classifiers to the FFHQ and FF datasets for A_t and A_p classification tasks like s, h, gl, a, g, and r. Our experimental analysis indicates that training the A_t classifier $\mathbb{D}$ on a limited balanced dataset can manifest biases related to A_p.

To mitigate this, $\mathbb{D}$ is trained on a composite dataset comprising the original balanced set and an augmented set derived via a GMM-based diffusion model. The performance of $\mathbb{D}$ is then evaluated on test subsets from both FFHQ and FF for A_t task accuracy and bias, using the metrics from Section 3.5. The results show that a classifier trained with the combined datasets exhibits lesser bias compared to one trained exclusively on the original dataset.

Fig. 2: An exemplar set of face images generated using DDPM. **Top:** Generated from high quality FFHQ dataset. **Bottom:** Generated from FairFace dataset.

4.3 Generating Faces with GMM and Diffusion Models

To synthesize a balanced dataset, we initially utilize the diffusion model's forward process via the HuggingFace Diffusers library. Subsequently, we implement a GMM on the resultant data, identifying the optimal component count based on BIC, which indicated that four components ($K = 4$) were ideal. The GMM utilizes the latent Gaussian distributions within the data for effective clustering. Post-clustering, we sample uniformly across these components to maintain class balance. For example, to generate 4,000 images across four components, we extract 1,000 samples from each. This approach is uniformly applied to both the FFHQ and FF datasets, leading to the stratification of facial images into distinct clusters that showcase prominent features.

4.4 Implementation Details

We use off-the-shelf classifiers with a ResNet-101 backbone to predict the classes in different downstream classification problems with binary targets (a, g, s, gl) and three-class targets (r, h). We optimize the training with the Adam optimizer and a learning rate of 10^{-4}. For the diffusion models, we train the Hugging Face Diffusers for 50 epochs also using the Adam optimizer with a learning rate of 10^{-4}. For the GMM model, we fit K=4 components for the latent code obtained from BIC to separate out our classes. We then use these means as

the means of the noise tensors during the reverse diffusion process to generate images. We set the covariance type in GMM model to be 'diagonal' and we generate an equal number of images from the bias corrected diffusion model. For a, g, s and gl classification, we use 5000 original samples per class and $5000/K$ generated images from each component. For r and h classification, we use 7000 original samples per class and $7000/K$ generated images from each component. We retrain our A_t classifiers with original training images and labels along with generated images and pseudolabels from the pretrained A_t classifier, and then perform evaluation on the original testing set.

5 Results and Discussion

We visualize an exemplar set of images generated unconditionally using DDPM in Figure 2 from both FFHQ and FF datasets. Tables 1 and 2 present a comparative analyses among a Baseline facial attribute classification model (classifier trained on only original dataset as defined before), a few existing debiasing techniques [8, 34, 50], and GAMMA-FACE across various A_p for A_t classification.

	FairFace								
	$A_t = g \mid A_p = a, r$			$A_t = r \mid A_p = a, g$			$A_t = a \mid A_p = r, g$		
Method	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)
[34]	0.187	1.36	81.67	0.237	**1.27**	78.10	0.112	1.502	78.62
[50]	0.142	1.38	84.14	0.163	1.393	79.3	**0.097**	**1.43**	80.13
[8]	0.169	1.53	82.28	0.218	1.781	75.5	0.130	1.62	76.51
Ours	**0.088**	**1.29**	**86.5**	**0.102**	1.36	**80.23**	0.128	1.510	**81.00**
	FFHQ								
	$A_t = s \mid A_p = a, g$			$A_t = h \mid A_p = a, g$			$A_t = gl \mid A_p = a, g$		
Method	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)
[34]	0.015	**1.48**	91.68	0.221	1.787	84.03	0.028	0.995	96.50
[50]	0.0064	1.61	93.16	0.153	1.798	**88.87**	0.031	1.008	97.29
[8]	0.019	1.77	91.58	0.192	1.84	82.11	0.040	1.156	96.10
Ours	**0.0056**	1.52	**94.84**	**0.146**	**1.756**	82.81	**0.0208**	**0.987**	**98.70**

Table 1: Quantitative assessment of GAMMA-FACE and existing debiasing techniques on FairFace and FFHQ dataset using bias evaluation metrics: Bias (B), Bias Amplification (BA) and Overall accuracy (Acc.). The up-arrow (↑) indicates higher Acc. while the down-arrow (↓) indicates lower B and BA values are preferable.

The Bias metric can be intuitively thought to measure the degree of disparity in the classification accuracy across different groups for A_p. GAMMA-FACE demonstrates a consistent reduction in bias, with the most significant decrease seen in g and and r classification tasks for the FF dataset and all three A_t classification tasks in the FFHQ dataset. It thus indicates an effective reduction in discriminatory decision-making patterns. The BA metric suggests that lower values are preferable, as they indicate a lessened association between A_p and A_t.

	FairFace					
	$A_t = g \mid A_p = a, r$		$A_t = r \mid A_p = a, g$		$A_t = a \mid A_p = r, g$	
Method	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)
Baseline	0	0.886	0	0.798	0	**0.769**
Ours	**0.085**	**0.801**	**0.118**	**0.740**	**0.454**	0.783
	FFHQ					
	$A_t = s \mid A_p = a, g$		$A_t = h \mid A_p = a, g$		$A_t = gl \mid A_p = a, g$	
Method	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)
Baseline	0	0.782	0	0.95	0	1.814
Ours	**0.673**	**0.698**	**0.4244**	**0.912**	**0.128**	**0.918**

Table 2: Quantitative assessment of GAMMA-FACE and existing debiasing techniques on FairFace and FFHQ dataset using bias evaluation metrics: Bias Performance Coefficient (BPC) and KL divergence (KL). The up-arrow (↑) indicates higher BPC while the down-arrow (↓) indicates lower KL values are preferable.

GAMMA-FACE notably reduces BA values to 1.29 for g classification task in the FF dataset, while taking values of 1.756 and 0.987 in the h and gl classification tasks for the FFHQ dataset respectively. An important consideration to make while choosing a bias mitigation technique is that we do not compromise on the overall performance of the model while reducing bias. Since our method involves data augmentation using diffusion generated de-biased data, we expect not to compromise on overall accuracy while reducing bias which is consistent across our experiments. We further compare our method to the Baseline model with two additional metrics- BPC and KL. In reviewing the effectiveness of GAMMA-FACE with respect to BPC, it reveals the capacity of our technique to enhance fairness while maintaining or improving classifier performance. The Baseline model exhibits a BPC of zero for all tasks (since Bias = $\text{Bias}_{deb}^{A_p}$ and TPR = $\text{TPR}_{deb}^{A_p}$). GAMMA-FACE shows a marked increase in BPC, such as 0.673 for s and 0.454 for g classification tasks in FFHQ and FF respectively, which illustrate that not only does GAMMA-FACE significantly reduce bias, but it also improves the overall TPR for the respective tasks. The KL divergence quantifies the difference in classifier score distributions, providing insight into the classifier's ability to maintain equalized odds. The baseline shows higher KL divergence values, indicating greater disparity and thus, a higher level of bias. GAMMA-FACE demonstrates a consistent reduction in KL divergence across all tasks. For example, in the gl classification task, KL divergence decreases from 1.814 to 0.918 in GAMMA-FACE, illustrating an improvement towards achieving equalized odds.

6 Ablation Studies

In previous experiments associated with an A_t, the training data for the sensitive attribute classifier consists of 50% original data and 50% diffusion-generated data to correct the bias. In this context, it is interesting to explore the effect of different ratios of original and diffusion generated data to train the classifier

	FairFace						FFHQ					
	$A_t = g \mid A_p = a, r$			$A_t = r \mid A_p = a, g$			$A_t = s \mid A_p = a, g$			$A_t = h \mid A_p = a, g$		
%Gen+%Org	B	BA	Acc.	B	BA	Acc.	B	BA	Acc.	B	BA	Acc.
100	0.117	1.59	83.25	0.125	1.72	72.83	0.0204	1.928	93.16	0.181	1.98	76.58
70+30	0.1098	1.41	82.25	**0.0956**	1.541	71.30	0.119	1.75	92.18	0.216	1.824	77.98
30+70	**0.089**	**1.33**	**86.21**	0.104	**1.354**	**77.93**	**0.0044**	**1.513**	**93.85**	**0.152**	**1.76**	**81.19**

Table 3: Ablation study on the effect of different mixing ratios (Generated + Original) in GAMMA-FACE on FairFace and FFHQ dataset evaluated using bias evaluation metrics: Bias (B), Bias Amplification (BA) and Overall accuracy (Acc.)

	FairFace				FFHQ			
	$A_t = g \mid A_p = a, r$		$A_t = r \mid A_p = a, g$		$A_t = s \mid A_p = a, g$		$A_t = h \mid A_p = a, g$	
%Gen+%Org	BPC	KL	BPC	KL	BPC	KL	BPC	KL
100	-0.266	1.01	-0.093	0.989	-0.243	0.95	-0.176	1.27
70+30	-0.204	1.23	-0.031	**0.85**	-0.201	0.852	-0.376	1.14
30+70	**0.117**	**0.978**	**0.055**	0.913	**0.0056**	**0.787**	**-0.0023**	**1.05**

Table 4: Ablation study on the effect of different mixing ratios (Generated + Original) in GAMMA-FACE on FairFace and FFHQ dataset using bias evaluation metrics: Bias Performance Coefficient (BPC) and KL divergence (KL).

for bias reduction. Tables 3 and 4 display the quantitative metrics for different mixing ratios of original and generated data for a binary and a three class classification task on FFHQ and FF respectively. It is observed that using only generated images to train and then testing on original dataset leads to severe degradation in performance because of a domain gap. However, as seen from the values, having a mixture ratio of 7:3 (Generated:Original) makes the performance a little bit better in comparison to when trained on 100% generated data. The best performance is indicated strongly by the BPC metric (trade-off between overall accuracy and bias) which is highest in the case of a 1:1 (Generated:Original) ratio across both datasets.

7 Conclusion

Accounting for biases in diffusion-based face generation models is essential to alleviate potential undesirable consequences that encoding protected attributes may have across diverse downstream target applications. In this work, we demonstrate a novel strategy, to counteract these biases using GMM-based disentanglement in the latent space of unconditional diffusion models. Our method offers an additional benefit in terms of performance as it is not employed in the training phase. Without retraining, we can effectively address bias in unconditional face generation by equally sampling from the GMM separated components and utilizing pretrained attribute classifiers. This is an initial work in exploiting the Gaussian structure of the latent code of diffusion models for fair generation and downstream applications.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

1. Albiero, V., Zhang, K., Bowyer, K.W.: How does gender balance in training data affect face recognition accuracy? In: 2020 ieee international joint conference on biometrics (ijcb). pp. 1–10. IEEE (2020)
2. Albiero, V., Zhang, K., King, M.C., Bowyer, K.W.: Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. IEEE Transactions on Information Forensics and Security **17**, 127–137 (2021)
3. Antoniou, A., Storkey, A., Edwards, H.: Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340 (2017)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)
6. Calimeri, F., Marzullo, A., Stamile, C., Terracina, G.: Biomedical data augmentation using generative adversarial neural networks. In: International conference on artificial neural networks. pp. 626–634. Springer (2017)
7. Choi, K., Grover, A., Singh, T., Shu, R., Ermon, S.: Fair generative modeling via weak supervision. In: International Conference on Machine Learning. pp. 1887–1898. PMLR (2020)
8. Dhar, P., Gleason, J., Roy, A., Castillo, C.D., Chellappa, R.: Pass: protected attribute suppression system for mitigating bias in face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15087–15096 (2021)
9. Dhar, P., Gleason, J., Roy, A., Castillo, C.D., Phillips, P.J., Chellappa, R.: Distill and de-bias: Mitigating bias in face verification using knowledge distillation. arXiv preprint arXiv:2112.09786 (2021)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
11. Duarte, A.C., Roldan, F., Tubau, M., Escur, J., Pascual, S., Salvador, A., Mohedano, E., McGuinness, K., Torres, J., Giro-i Nieto, X.: Wav2pix: Speech-conditioned face generation using generative adversarial networks. In: ICASSP. pp. 8633–8637 (2019)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

13. Grover, A., Song, J., Kapoor, A., Tran, K., Agarwal, A., Horvitz, E.J., Ermon, S.: Bias correction of learned generative models using likelihood-free importance weighting. Advances in neural information processing systems **32** (2019)
14. Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., Wang, J.: Long text generation via adversarial training with leaked information. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Huang, Z., Chan, K.C., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6080–6090 (2023)
17. Jain, N., Olmo, A., Sengupta, S., Manikonda, L., Kambhampati, S.: Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat face lenses. Artificial Intelligence **304**, 103652 (2022)
18. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1548–1558 (2021)
19. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
22. Kashyap, R.: Inconsistency of the aic rule for estimating the order of autoregressive models. IEEE Transactions on Automatic Control **25**(5), 996–998 (1980)
23. Krishnapriya, K., Albiero, V., Vangara, K., King, M.C., Bowyer, K.W.: Issues related to face recognition accuracy varying based on race and skin tone. IEEE Transactions on Technology and Society **1**(1), 8–20 (2020)
24. Lee, J., Kim, E., Lee, J., Lee, J., Choo, J.: Learning debiased representation via disentangled feature augmentation. Advances in Neural Information Processing Systems **34**, 25123–25133 (2021)
25. Liu, M.Y., Huang, X., Yu, J., Wang, T.C., Mallya, A.: Generative adversarial networks for image and video synthesis: Algorithms and applications. Proceedings of the IEEE **109**(5), 839–862 (2021)
26. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408 (2023)
27. Maluleke, V.H., Thakkar, N., Brooks, T., Weber, E., Darrell, T., Efros, A.A., Kanazawa, A., Guillory, D.: Studying bias in gans through the lens of race. In: European Conference on Computer Vision. pp. 344–360. Springer (2022)
28. Michael, K., Abbas, R., Jayashree, P., Bandara, R.J., Aloudat, A.: Biometrics and ai bias. IEEE Transactions on Technology and Society **3**(1), 2–8 (2022)
29. Mittermaier, M., Raza, M.M., Kvedar, J.C.: Bias in ai-based models for medical applications: challenges and mitigation strategies. npj Digital Medicine **6**(1), 113 (2023)
30. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., et al.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Scientific Reports **13**(1), 12098 (2023)

31. Naik, R., Nushi, B.: Social biases through the text-to-image generation lens. arXiv preprint arXiv:2304.06034 (2023)
32. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)
33. Perera, M.V., Patel, V.M.: Analyzing bias in diffusion-based face generation models. In: 2023 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2023)
34. Ramaswamy, V.V., Kim, S.S., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9301–9310 (2021)
35. Reynolds, D.A., et al.: Gaussian mixture models. Encyclopedia of biometrics **741**(659-663) (2009)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
37. de Rosa, G.H., Papa, J.P.: A survey on text generation using generative adversarial networks. Pattern Recognition **119**, 108098 (2021)
38. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
39. Schwarz, G.: Estimating the dimension of a model. The annals of statistics pp. 461–464 (1978)
40. Sharmanska, V., Hendricks, L.A., Darrell, T., Quadrianto, N.: Contrastive examples for addressing the tyranny of the majority. arXiv preprint arXiv:2004.06524 (2020)
41. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
42. Singh, J., Gould, S., Zheng, L.: High-fidelity guided image synthesis with latent diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5997–6006. IEEE (2023)
43. Stypułkowski, M., Vougioukas, K., He, S., Zięba, M., Petridis, S., Pantic, M.: Diffused heads: Diffusion models beat gans on talking-face generation. arXiv preprint arXiv:2301.03396 (2023)
44. Tan, S., Shen, Y., Zhou, B.: Improving the fairness of deep generative models without retraining. arXiv preprint arXiv:2012.04842 (2020)
45. Teo, C.T., Abdollahzadeh, M., Cheung, N.M.: Fair generative models via transfer learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2429–2437 (2023)
46. Wang, A., Russakovsky, O.: Directional bias amplification. In: International Conference on Machine Learning. pp. 10882–10893. PMLR (2021)
47. Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequity in object detection. arXiv preprint arXiv:1902.11097 (2019)
48. Wu, H., Albiero, V., Krishnapriya, K., King, M.C., Bowyer, K.W.: Face recognition accuracy across demographics: Shining a light into the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1041–1050 (2023)
49. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys **56**(4), 1–39 (2023)

50. Zhang, F., He, Q., Kuang, K., Liu, J., Chen, L., Wu, C., Xiao, J., Zhang, H.: Distributionally generative augmentation for fair facial attribute classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22797–22808 (2024)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)