Improving Fairness using Vision-Language Driven Image Augmentation

Moreno D'Incà[1] Christos Tzelepis[2] Ioannis Patras[2] Nicu Sebe[1]

[1]University of Trento
moreno.dinca@unitn.it, niculae.sebe@unitn.it

[2]Queen Mary University of London
{c.tzelepis, i.patras}@qmul.ac.uk

Abstract

Fairness is crucial when training a deep-learning discriminative model, especially in the facial domain. Models tend to correlate specific characteristics (such as age and skin color) with unrelated attributes (downstream tasks), resulting in biases which do not correspond to reality. It is common knowledge that these correlations are present in the data and are then transferred to the models during training (e.g., [35]). This paper proposes a method to mitigate these correlations to improve fairness. To do so, we learn interpretable and meaningful paths lying in the semantic space of a pre-trained diffusion model (DiffAE) [27] – such paths being supervised by contrastive text dipoles. That is, we learn to edit protected characteristics (age and skin color). These paths are then applied to augment images to improve the fairness of a given dataset. We test the proposed method on CelebA-HQ and UTKFace on several downstream tasks with age and skin color as protected characteristics. As a proxy for fairness, we compute the difference in accuracy with respect to the protected characteristics. Quantitative results show how the augmented images help the model improve the overall accuracy, the aforementioned metric, and the disparity of equal opportunity. Code is available at: `https://github.com/Moreno98/Vision-Language-Bias-Control`.

1. Introduction

Today's society is careful about ethical topics and with the raising of publicly available AI tools [16, 29, 30] concerns about their fairness are also growing. In a supervised learning setting, the importance of the training data is well-known since the behavior of the model at inference time is highly correlated to the seen data. Modern models can effectively learn and highly perform multiple downstream tasks generalizing to unseen data. Besides the effectiveness of the pipeline, training data also brings unwanted side effects. It has been proven that vision datasets contain biases [35], thus the models learn the correlations present in the data which may be malignant [4, 7, 13, 43]. These concerns become a particularly sensitive subject when it comes to the facial domain. Modern machine learning models are dominant at a wide range of applications, such as face/emotion recognition and mask detection [11, 16, 20]. In this context, studying the behavior of deep learning models is crucial to avoid unwanted situations at inference time [26]. For example, the model's performance may drop when presented with a particular protected characteristic (e.g., very young/old or dark-skinned faces). The above issues motivate us to study the behavior of a deep learning model with respect to facial protected characteristics which are sensitive to society and can raise ethical concerns.

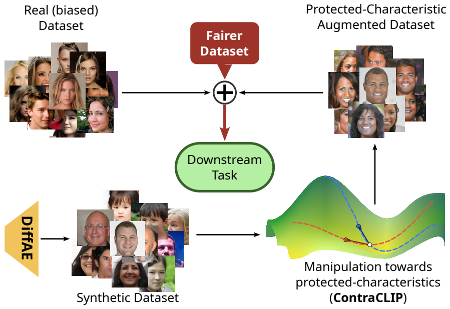


Figure 1. We make a biased dataset fairer by augmenting it with images (generated by DiffAE [27]) depicting the desired protected characteristics (e.g., dark-skinned people) after being manipulated by our ContraCLIP [36]-based text-driven augmentation module.

Training fair discriminative models has become of paramount importance for the research community during the past years. Recent works have shown that not only do models learn the underlying bias present in the data [14, 34], but they tend to often amplify it [13, 38, 43]. Multiple techniques have been proposed for mitigating the bias, from task-specific training, such as the introduction of regularization terms or architectural approaches [22, 32, 38] to data augmentation strategies [2, 18].

Recently, generative models, such as Generative Adversarial Networks (GANs) [12], have shown remarkable performance in a multitude of tasks through discovering controllable generative paths in their latent or feature spaces [5, 6, 23, 24, 37]. Thus, GAN-based methods have been employed as a data augmentation technique to generate fairer data [8, 39, 40], to generate counterfactuals [1, 9] or to generate counterparts by editing sensitive attributes [41]. The above works train generative models from scratch which may be impractical, especially in low data regimes. Additionally, the pre-trained generative models are expected to reflect the bias that is inherent to the datasets where they have been trained on [25, 39, 41], challenging those methods that use them for bias mitigation.

In this paper, we address the above limitations by proposing a novel approach that leverages a *pre-trained* diffusion model [27, 33] to edit sensitive attributes in facial images, in order to improve the fairness of existing (biased) datasets and, consequently, the fairness of a discriminative model trained on such datasets. By contrast to previous works that train generative models (e.g., GANs [12]) from scratch [9, 39, 41], we incorporate the power of a fixed pre-trained diffusion model to change sensitive facial attributes from a pool of generated images. The manipulated faces (with respect to the desired sensitive attributes) are used to make the original dataset fairer and mitigate the bias present on a downstream model trained on the original dataset. We illustrate this in Fig. 1. Our setting consists of a binary downstream classification attribute and a binary sensitive attribute towards which the model may exhibit bias. Throughout the paper, we will be referring to the downstream classification attribute as *attribute* and to the sensitive attribute as *protected characteristic*. The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to leverage diffusion models and a vision-language model in the context of fairness on discriminative models.
- We propose to edit the protected characteristics by learning interpretable paths in the semantic space of DiffAE [27] guided by natural language without the need to fine-tune the *pre-trained* generative model.
- We test our method on several downstream tasks of CelebA-HQ [21] and UTKFace [42] with *skin color* and *age* as protected characteristics showing competitive or better performance when mitigating the bias.
- We show that our method is capable of increasing the bias towards a specific attribute if needed – that is, in contrast to previous works (e.g., [22, 38, 41]), our method can control (i.e., decrease or increase) the bias concerning a specific attribute.

2. Related work

During recent years, the research community has directed its efforts towards mitigating bias and improving fairness in discriminative models mainly adopting one of the following approaches: i) proposing training techniques without changing the training data at hand or ii) applying some sort of data augmentation for the under-represented classes. We review each category below.

Training techniques Different strategies have been investigated when mitigating the bias at training time, from the employment of regularization terms in the loss [22] to architectural methods [32, 38]. Wang et al. [38] proposed a new benchmark for bias mitigation by studying various techniques, such as oversampling rare examples, adversarial training and domain discriminative training, and proposing independent domain training where the downstream task is independently learnt by leveraging two independent heads, one for each domain class, leading to a model that is aware of the sensitive domain. Nam et al. [22] proposed Learning from Failure (LfF), which simultaneously trains two classifiers, one to be biased and the other to be unbiased, by focusing on the hard samples for the biased one. This setting works under the assumption that a malignant bias is learnt when the sensitive attribute is easier to learn than the downstream target attribute; thus, if a biased model is struggling in learning a set of samples, then those samples will help in unbiasing a second model. This is achieved via generalized cross-entropy for amplifying the bias on the first model and via a weighted cross-entropy loss with relative difficulty for the target unbiased model. Savani et al. [32] introduce three intra-processing methods for bias mitigation, namely random perturbation where at each training iteration the model's weights are multiplied by Gaussian noise, layer-wise optimization where each layer of the network is optimized separately with a common objective function, and adversarial training where the model's bias is predicted via a trainable critic and used to improve the fairness avoiding the non-differentiable issue of bias metrics.

Generative data augmentation A certain line of research proposes the generation of "fairer" data using generative models towards improving fairness. One of the first works in this direction is [39], where Sattigeri et al. introduce Fairness GAN, a GAN [12] conditioned on the protected sensitive attribute. This work generates a fairer dataset by training the proposed architecture on the original dataset. Dash et al. [9] introduced a model trained to generate counterfactual versions of the same image based on the knowledge from a pre-defined causal graph. The synthesized images are then used to mitigate the bias by adding a regularization term to the loss, minimizing the MSE between the logits of the model when presented by the original image and the counterfactual. A drawback of this work is that the prior knowledge that is required from the causal graph for encoding the attribute and protected characteristic relations may not be readily available in practice. Zhang et al. [41] proposed a new setting where sensitive labels and downstream attributes are partially annotated. This work

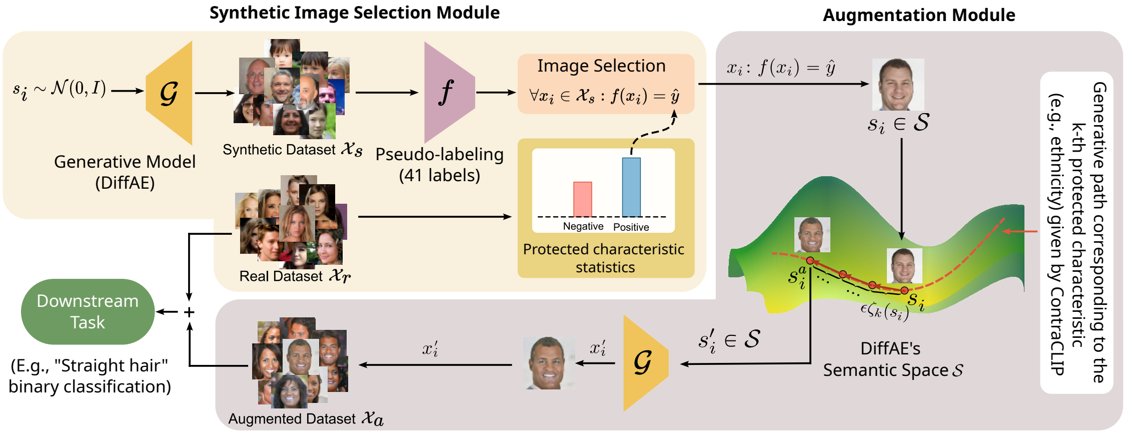

Synthetic Image Selection Module
Augmentation Module
$s_i \sim \mathcal{N}(0, I)$
$\mathcal{G}$
Generative Model (DiffAE)
Synthetic Dataset $\mathcal{X}_s$
f
Pseudo-labeling (41 labels)
Image Selection
$\forall x_i \in \mathcal{X}_s : f(x_i) = \hat{y}$
Negative
Positive
Protected characteristic statistics
Real Dataset $\mathcal{X}_r$
Downstream Task
+
(E.g., "Straight hair" binary classification)
$x_i : f(x_i) = \hat{y}$
$s_i \in \mathcal{S}$
s_i^a
s_i
$\epsilon\zeta_k(s_i)$
Generative path corresponding to the k-th protected characteristic (e.g., ethnicity) given by ContraCLIP
DiffAE's Semantic Space $\mathcal{S}$
$s_i' \in \mathcal{S}$
$\mathcal{G}$
x_i'
x_i'
Augmented Dataset $\mathcal{X}_a$

Figure 2. Overview of the proposed Vision-Language Bias Control (VLBC) method for controlling the bias in facial image datasets. Given a real training set $\mathcal{X}_r$ and a downstream task, we find the under-represented protected characteristic (e.g., black in *skin colour*) by computing the sample statistics. Based on this, we select which images from a synthetic dataset $\mathcal{X}_s$ (generated using the DiffAE [27] generator $\mathcal{G}$ and pseudo-labelled by a pre-trained network on 41 attributes f [15,21]) to use for augmentation. Then, the selected images are manipulated by our augmentation module (ContraCLIP+DiffAE), pre-trained on text prompts defining the desired protected characteristic. In this example, we manipulate/augment the selected images based on *skin colour*. Note that the original labels of the augmented images (i.e., corresponding to the attribute class at hand) do not change. Finally, the augmented dataset $\mathcal{X}_a$ is used along with the original real dataset $\mathcal{X}_r$ for training downstream tasks.

first learns a generator and a classifier with the available annotations, while subsequently the two models are incrementally trained using the mutual outputs. This semi-supervised setting of [41] is necessary due to the lack of annotations needed by the generator during training. The study also introduces a contrastive learning framework with balanced augmented data; i.e., for each original image, its counterpart is generated by changing the sensitive attribute, and then the downstream model is presented with both versions of the image pushing together images with different sensitive attributes and pushing away images with same sensitive attribute. Similarly, Jung et al. [14] proposed the training of a model to pseudo-label a dataset with a specific sensitive attribute. Then, the pseudo-labels are filtered and replaced with random choices if the prediction confidence is lower than a certain threshold. Finally, the labeled and pseudo-labeled data are used to train a fairer model employing a training technique for bias mitigation (e.g., FairHSIC [28]).

The above generative techniques achieve notable bias mitigation results, but they typically require the training of domain-related generative models, demanding a significant amount of domain-specific data that may not be available for particular sensitive attributes because of their scarcity during data collection. Other works mitigate this problem via semi-supervised training [41]. In this work, we argue that a better solution exists. Since training a generative model on the available data may lead to a biased generator which could make the generation of rare examples challenging [39], we propose to explicitly learn to edit sensitive attributes by exploiting a *pre-trained* generative model without the need of training or fine-tuning it.

3. Vision-Language Bias Control (VLBC)

In this section, we present our method for controlling the bias of a given dataset via augmenting images by editing specific protected characteristics (e.g., *skin colour*) using a diffusion-based generative model (DiffAE [27]) and an augmentation module that learns to generate images driven by prompts in natural language (ContraCLIP [36]). An overview of the proposed framework, which we call Vision-Language Bias Control (VLBC), is shown in Fig. 2. Concretely, given a real training dataset $\mathcal{X}_r$ of facial images, annotated for several attribute classes (e.g., CelebA-HQ's [21] attributes, such as *chubby*, *long hair*, etc.), we first calculate, for each class, the number of positive and negative samples with respect to a protected characteristic (e.g., *skin colour*). By doing so, we identify whether the protected characteristic at hand is under-represented in the given dataset. Then, after having identified the bias towards a specific protected characteristic, we may control it (i.e., mitigate or increase it) by i) selecting fake images from a large datasets of synthetic facial images generated by DiffAE [27], $\mathcal{X}_s$, that have been pseudo-labelled by a pre-trained network on 41 attributes f [15,21] (*Synthetic Image Selection*) and ii) manipulating accordingly using the proposed *Augmentation Module*. The augmented dataset, $\mathcal{X}_a$, is then used along with the original (biased) dataset $\mathcal{X}_r$ towards training fairer downstream classifiers. We note that we do not merely sample synthetic images from $\mathcal{X}_s$ since we do not possess any control over the attributes of the generated images and there is no guarantee that generated images will be numerically adequate to compensate for the under-

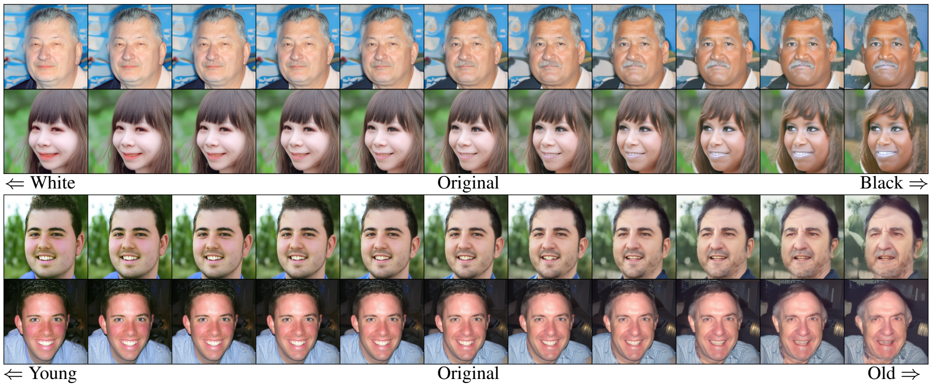


Figure 3. Qualitative manipulation results on *skin color* and *age* (protected characteristics) using the proposed ContraCLIP+DiffAE.

represented classes of protected characteristics. This is simply because synthetic images follow the dataset distribution where the generative models have been trained, thus, they still suffer from biases that are present in those datasets. We prove this intuition by reporting the failing of this method on *skin color* (see *Baseline-sampling* Sect.4.1 and Table 3).

3.1. Synthetic image selection module

In this section we present the *Synthetic Image Selection Module* of our framework (see top-left part of Fig. 2). As discussed above, given a real dataset $\mathcal{X}_r$, our goal towards mitigating the bias with respect to a specific protected characteristic is to augment the dataset with images that exemplify that protected characteristic ($\mathcal{X}_a$), so as the resulting dataset ($\mathcal{X}_r + \mathcal{X}_a$) is fairer in a given downstream task.

The aforementioned images that will be manipulated to exemplify the desired protected characteristic are selected from a synthetic dataset $\mathcal{X}_s$ of images generated by the DiffAE [27]. The images of $\mathcal{X}_s$ are pseudo-labelled by a network f pre-trained on the CelebA's [21] 40 facial attributes (40 binary pseudo-labels) and on FairFace's [15] *skin colour* prediction (1 label). We note that while FairFace [15] provides predictions with respect to *skin colour* in terms of four groups (i.e., "white", "black", "Asian", and "Indian"), we use a binary label corresponding to *white/black* by removing data from the "Indian" group. The resulting annotated synthetic dataset $\mathcal{X}_s$ is given as $\mathcal{X}_s = \{(x_i, s_i, \hat{y}_i)\}_{i=1}^{N_s}$, where N_s denotes the dataset size (in our experiments we set $N_s = 120,000$), x_i denotes the i-th image and s_i its code in the semantic space of DiffAE [27].

Finally, in order to decide on the sort of images required for augmentation (image selection), we calculate the statistics over the real (training) set $\mathcal{X}_r$ by taking into account both protected characteristics and classification attributes. Concretely, we calculate the number of samples that are annotated for the desired classification attribute and that appear as positive or negative with respect to the desired protected characteristic. In the case of *bias mitigation* (towards reducing the bias), we increase the number of images of the minority protected characteristic class, keeping unaltered the downstream attribute labels, by selecting synthetic images x_i for which the pseudo-label $f(x_i)$ is the majority class (e.g., white people see Fig. 2) – we denote this variant VLBC-. On the contrary, in the case of *bias amplification* (towards increasing the bias), we select to manipulate synthetic images pseudo-labeled with the minority class for the protected characteristic class at hand – we denote this variant of our method VLBC+. We note that bias mitigation is typically the desired behaviour of the resulting augmented dataset, however our formulation allows for increasing the bias as well, which is designed for research purposes only, towards studying bias in real datasets and downstream tasks. After selecting which images to manipulate towards the desired protected characteristic, we apply the proposed augmentation module, which we describe below.

Semantic dipoles text-prompts		
Negative direction $(-)$		**Positive direction** $(+)$
Young	↔	Old
"An ID photo of a young person."	↔	"An ID photo of an old person."
White skin color	↔	Black skin color
"A pale skin face."	↔	"A black face."

Table 1. Text prompts used to learn the described manipulations.

3.2. Augmentation module (ContraCLIP+DiffAE)

In this section, we present the proposed augmentation module that is capable of manipulating facial images with respect to a protected characteristic (e.g., *skin color*) described in natural language. For doing so, we build on the work of Tzelepis et al. [36] and we modify ContraCLIP so that it discovers language-driven controllable paths in the semantic space of DiffAE [27], instead of that of StyleGAN2 [17]. We briefly discuss both components below.

Diffusion AutoEncoder (DiffAE) [27] has been recently proposed to endow a Denoising Diffusion Implicit Model (DDIM) [33] with a meaningful semantic space $\mathcal{S}$ and, thus,

Task	Method	Accuracy ↑	f1-score ↑	Acc Diff	Fairness	
					Δ_A ↓	Δ_M ↓
Wearing Necktie	Baseline	94.37±0.11	78.19±0.24	10.65±0.31	6.02±0.27	6.59±0.54
	Baseline-sampling	94.29±0.11	77.46±0.59	10.97±0.26	4.59±1.57	6.19±2.13
	Weighting [38]	94.46±0.04	75.14±0.09	11.01±0.06	5.88±1.14	9.99±2.43
	LfF [22]	**94.84±0.13**	**81.11±0.46**	10.30±0.30	**4.11±0.60**	6.35±0.06
	CGL-FairHSIC [14]	92.83±0.00	48.14±0.00	15.88±0.00	–	–
	VLBC- (ours)	94.69±0.06	78.48±0.30	**10.07±0.13**	5.18±0.35	5.83±0.75
	VLBC- \f (ours)	94.65±0.04	78.43±0.21	10.18±0.08	5.21±0.32	**5.72±0.47**
Chubby	Baseline	93.24±0.04	72.56±0.88	15.85±0.65	8.24±1.42	9.65±1.92
	Baseline-sampling	93.16±0.15	71.6±0.39	**15.51±0.48**	**3.58±1.1**	**5.22±0.53**
	Weighting [38]	93.33±0.09	66.33±0.72	16.75±0.21	9.64±0.57	18.15±1.38
	LfF [22]	92.60±0.51	**76.74±0.74**	15.76±0.94	18.28±2.65	21.77±3.61
	CGL-FairHSIC [14]	92.49±0.01	48.93±0.82	19.41±0.10	–	–
	VLBC- (ours)	**93.44±0.02**	71.96±0.45	15.69±0.22	5.39±0.37	5.79±0.28
	VLBC- \f (ours)	93.4±0.06	71.91±0.28	15.88±0.2	4.4±0.35	5.86±0.06
Arched Eyebrows	Baseline	80.15±0.3	78.91±0.34	-9.00±0.20	9.22±0.12	12.41±0.36
	Baseline-sampling	80.18±0.22	79.15±0.22	-8.25±0.17	9.73±0.22	12.75±0.22
	Weighting [38]	79.57±0.13	78.29±0.14	-9.18±0.29	**6.25±0.26**	**9.01±0.45**
	LfF [22]	25.27±0.21	25.06±0.21	9.36±0.50	–	–
	CGL-FairHSIC [14]	**84.17±0.46**	**83.28±0.46**	**-7.63±0.48**	6.45±1.38	10.35±1.28
	VLBC- (ours)	80.44±0.06	79.27±0.04	-8.62±0.19	10.56±0.20	13.00±0.53
	VLBC- \f (ours)	80.56±0.14	79.39±0.14	-8.47±0.28	10.39±0.04	12.75±0.55
Double Chin	Baseline	94.17±0.13	74.47±0.30	15.15±0.6	18.74±1.38	27.86±2.94
	Baseline-sampling	94.36±0.23	74.21±0.58	**13.71±0.85**	10.87±2.25	15.4±4.51
	Weighting [38]	**94.66±0.09**	69.09±0.78	14.46±0.25	**1.69±0.34**	**2.00±0.24**
	LfF [22]	94.16±0.08	**78.39±0.29**	14.27±0.13	21.94±1.01	30.54±1.89
	CGL-FairHSIC [14]	93.74±0.00	48.39±0.00	18.67±0.00	–	–
	VLBC- (ours)	94.48±0.04	74.35±0.23	14.72±0.21	14.20±0.07	20.69±0.18
	VLBC- \f (ours)	94.46±0.04	74.19±0.33	14.57±0.15	15.01±0.25	22.44±0.48

Table 2. Results of the classification tasks with **age** as protected characteristic. We train the model with the original training data (baseline), with the original data plus synthetic (baseline-sampling), and with the proposed VLBC-. We compare with *weighting* [38], LfF [22], and CGL-FairHSIC [14].

the ability of semantic editing.

ContraCLIP [36] Given a pre-trained GAN, ContraCLIP [36] learns non-linear paths in its latent space driven by contrasting semantic dipoles given in natural language. To do so, one needs to define pairs of sentences that convey contrasting meanings and express the limits of the interpretation that are required by the optimized latent paths to encode. Each such pair corresponds to one trainable path in the GAN latent space. Given K semantic dipoles, ContraCLIP first represents them in the CLIP text space and then use them as the centres of K RBF-based warping functions [36] $\{\zeta_k\}_{k=1}^K$. This gives rise to K vector fields, which provide paths from the one pole/sentence to the other and can be used as the supervisory signal that guides the trainable paths, for any given image and its transformed version along a certain latent path.

In this work, we introduce our augmentation module (illustrated in the bottom-right part of Fig. 2) by building on ContraCLIP [36] and extending it towards learning generative paths in the semantic space $\mathcal{S}$ of DiffAE [27]. We remark that, in our pipeline, we do not train the DiffAE's [27] generator $\mathcal{G}$ (that is *pre-trained* on FFHQ [16]). Hereafter, we will be referring to our augmentation module as ContraCLIP+DiffAE. We pre-train ContraCLIP+DiffAE to learn paths that refer to protected characteristics (see Tab. 1). Then, given the semantic code $s_i \in \mathcal{S}$ of the i-th synthetic image, we can manipulate it by traversing across the k-th path (that corresponds to the k-th protected characteristic; e.g., *skin colour*) by performing steps given by $s'_i = s_i + \epsilon\zeta_k(s_i)$, where ζ_k is the warping function [36] for the k-th protected characteristic and ϵ is scalar determining

Task	Method	Accuracy ↑	f1-score ↑	Acc Diff	Fairness	
					Δ_A ↓	Δ_M ↓
Straight Hair	Baseline	82.52±0.53	71.88±0.74	10.17±0.81	20.16±3.61	32.85±6.24
	Baseline-sampling	81.96±0.26	72.12±0.27	10.57±0.34	22.17±2.9	35.38±5.38
	Weighting [38]	81.88±0.27	71.01±0.46	10.75±0.76	16.53±3.13	25.61±5.80
	LfF [22]	39.50±3.24	39.12±2.89	17.46±0.61	–	–
	CGL-FairHSIC [14]	**84.51±0.12**	**76.15±0.61**	**9.00±0.72**	15.93±2.48	**24.65±5.78**
	VLBC- (ours)	81.99±0.08	70.78±0.08	9.48±0.22	**15.56±0.81**	25.39±1.44
	VLBC- \f (ours)	82.03±0.14	70.63±0.3	9.44±0.14	15.75±0.33	25.91±0.65
Young	Baseline	85.25±0.09	79.43±0.26	-9.04±0.09	8.96±0.30	10.40±1.20
	Baseline-sampling	85.33±0.36	79.8±0.41	-10.95±0.8	10.2±0.74	11.85±0.5
	Weighting [38]	85.1±0.09	78.87±0.23	-7.23±1.86	8.15±2.66	11.96±4.44
	LfF [22]	21.18±0.54	21.16±0.56	5.51±1.27	–	–
	CGL-FairHSIC [14]	**87.93±0.28**	**82.25±0.82**	-7.57±1.26	7.27±1.65	10.03±2.45
	VLBC- (ours)	85.55±0.04	79.10±0.27	**-6.27±0.56**	**5.32±0.55**	**5.46±0.59**
	VLBC- \f (ours)	85.5±0.09	79.27±0.15	-7.38±0.48	5.68±0.29	7.52±0.75
Wearing Necktie	Baseline	94.35±0.11	77.49±0.20	-8.10±0.63	11.00±1.93	17.13±3.67
	Baseline-sampling	94.4±0.03	78.17±0.31	-7.86±0.6	11.43±0.94	18.25±1.82
	Weighting [38]	94.33±0.12	76.74±0.33	-5.48±0.27	10.66±1.15	19.89±2.30
	LfF [22]	**94.86±0.07**	**81.27±0.06**	-7.49±0.47	15.01±1.85	26.28±4.06
	CGL-FairHSIC [14]	93.65±0.78	61.18±10.09	**-3.88±0.41**	**0.32±0.22**	**0.5±0.28**
	VLBC- (ours)	94.45±0.04	76.81±0.32	-6.38±0.11	3.67±0.84	4.46±0.01
	VLBC- \f (ours)	94.46±0.04	76.77±0.19	-6.29±0.11	3.49±0.5	4.37±0.06
Big Lips	Baseline	65.62±0.54	57.62±0.12	6.32±1.47	45.17±2.41	54.23±0.82
	Baseline-sampling	**65.78±0.67**	**58.5±0.87**	8.94±0.96	51.82±1.21	58.33±0.54
	Weighting [38]	63.32±0.13	58.04±0.71	-0.84±3.48	**23.03±2.35**	**28.48±1.60**
	LfF [22]	35.45±0.22	34.58±0.41	-4.97±0.57	–	–
	CGL-FairHSIC [14]	65.72±1.04	58.29±0.93	**-0.71±2.48**	35.09±4.34	40.79±4.45
	VLBC- (ours)	65.69±0.04	57.67±0.14	5.77±0.63	43.74±0.42	53.05±0.77
	VLBC- \f (ours)	65.66±0.07	57.91±0.33	5.02±0.36	41.04±0.38	50.38±0.17
Big Nose	Baseline	78.82±0.07	74.74±0.21	3.53±0.35	29.72±1.96	31.88±2.59
	Baseline-sampling	**79.0±0.17**	**74.88±0.35**	6.9±0.52	35.45±0.63	36.71±0.46
	Weighting [38]	77.77±0.28	73.62±0.27	**1.29±0.89**	**18.71±0.69**	**23.12±0.77**
	LfF [22]	22.92±0.52	22.68±0.59	5.19±1.26	–	–
	CGL-FairHSIC [14]	78.59±0.22	74.36±0.41	7.73±0.3	36.51±1.36	37.92±1.03
	VLBC- (ours)	78.26±0.06	74.23±0.12	2.40±0.33	22.01±0.60	25.04±0.88
	VLBC- \f (ours)	78.37±0.1	74.32±0.03	2.48±0.18	23.41±0.49	25.78±0.21

Table 3. Results of the classification tasks with **skin color** as protected characteristic. We train the model with the original training data (baseline), with the original data plus synthetic (baseline-sampling), and the proposed VLBC-. We compare with *weighting* [38], LfF [22], and CGL-FairHSIC [14].

the length and the direction of the manipulation step.

It is worth noting that the traversal length in the semantic space $\mathcal{S}$ affects the degree of manipulation. Concretely, in order to enforce diversity during the augmentation phase and guarantee that manipulation is effective (i.e., it changes the characteristic at hand adequately), we define a minimum ($\mathcal{E}_{\min}$) and a maximum ($\mathcal{E}_{\max}$) number of traversal steps, and we randomly (uniformly) sample the number of steps ($\mathcal{E}$) in $[\mathcal{E}_{\min}, \mathcal{E}_{\max}]$ at each augmentation. That is, given a synthetic image $x_i \in \mathcal{X}_s$, we manipulate its semantic code by applying $\mathcal{E}$ steps before we arrive at the final augmented image $s_i^a \in \mathcal{S}$. As a result, after manipulating all selected synthetic images x_i, we obtain an augmented dataset $\mathcal{X}_a$, that is used along with the real dataset $\mathcal{X}_r$ in order to make it fairer with respect to the downstream task at hand. This is illustrated in the bottom-right part of Fig. 2.

4. Experiments

In this section, we present the experimental evaluation of the proposed framework for controlling the bias in facial datasets with respect to the protected characteristics *skin color* and *age*, towards the downstream task of binary attribute classification. We note that we train the classification models only on the classification attributes, not the protected characteristic. We provide qualitative and quantitative results on both mitigating (VLBC-, Sect. 4.1) and increasing (VLBC+, Sect. 4.2) the bias, and we compare with the state-of-the-art (SOTA) works of [14, 22, 38]. We first train a baseline model on the original training set, quantifying its initial bias without applying any fairness-related technique. Then we investigate how the same model behaves when fine-tuned on the augmented dataset created to mitigate or increase the bias. Moreover, while mitigating the bias, we introduce a second baseline, referred to as *baseline-sampling*, which injects synthetic images of the desired protected characteristic (e.g., black people) in the training set without applying the proposed augmentations. This baseline provides a simple yet effective way of determining whether and when our augmentation scheme is necessary, providing additional insight into the usefulness of synthetic images. As discussed in Sect. 3, we draw our intuition from the fact that generative model are inherently biased as well, since they depend on training on biased datasets. Finally, in Sect. 4.3, we present our ablation study.

Implementation details We evaluate our method using *MobileNetV2* [3,31] and we train our models on one Nvidia RTX A6000 with SGD and Focal Loss [19]. Learning rate is 10^{-3} when training from scratch (starting with ImageNet [10] weights) and 10^{-4} when fine-tuning. The baselines are trained for 100 epochs while fine-tuning (VLBC) is performed for 50 epochs for both *age* and *skin color*. We average the results over three runs and report mean/std.

Datasets We evaluate the proposed framework on (i) **CelebA-HQ** [21], a diverse dataset in terms of *skin colour*, *gender*, and *age*, which contains $30,000$ images annotated with 40 attributes, and (ii) **UTKFace** [42], an in-the-wild dataset consisting of $18,038$ face images annotated with respect to *gender*, *age* (spanning 116 years), and *skin colour*.

Evaluation metrics Under the binary classification setting, a natural way for describing fairness is to have a model performing *equally* regardless of the protected characteristic. For instance, predicting *big lips* should perform independently of characteristics such as *age* or *skin color*. Following this intuition, we calculate the accuracy of the downstream task conditioned on the protected characteristic [39]. E.g., in the case of the protected characteristic of *age*, we split the attribute classification accuracy into "young" and "old". We then calculate the difference between the two accuracies to capture the model's fairness. Ideally, a model will exhibit equal behavior on a zero-valued difference in accuracy. We also note that the sign of the difference in accuracy is indicative of the "direction" of the bias – i.e., a negative difference value would indicate a bias towards elder people, and vice versa for a positive value. We report the overall accuracy, the f1-score, and the difference in accuracy (Acc Diff). Additionally, we calculate the mean (Δ_A) and max (Δ_M) disparity of opportunity similarly to [14].

4.1. Bias mitigation with VLBC-

In this section, we show how the proposed framework for mitigating the bias (VLBC-) improves the fairness of a given dataset, assesed on a subset of the attribute classification tasks of CelebA-HQ [21]. We chose this based on the bias of a baseline model trained on all the CelebA-HQ attributes. That is, we ranked the tasks based on the difference in accuracy, and we chose the ones with higher values. Specifically, we decided to evaluate our method on the attribute classification tasks of *wearing necktie*, *chubby*, *arched eyebrows*, and *double chin* having *age* as the protected characteristic, and on *straight hair*, *young*, *wearing necktie*, *big lips*, and *big nose* having skin color as the protected characteristic. We employed our augmentation module (ContraCLIP+DiffAE) to balance the training set statistics with respect to the protected characteristics. For example, given big nose and skin color as pair attribute-protected characteristics, we sample images from $\mathcal{X}_s$ of white people (majority class) with and without *big nose* editing them into black people (minority class) balancing the training set.

Age				
Task	Method	Accuracy ↑	f1-score ↑	Acc Diff
Gender	Baseline	82.87±0.34	82.86±0.34	4.8±1.72
	Baseline-sampling	83.85±0.43	83.84±0.43	4.83±0.71
	Weighting [38]	83.78±0.12	83.77±0.13	6.17±0.68
	LfF [22]	44.72±3.19	43.61±3.43	-2.9±0.93
	CGL-FairHSIC [14]	**91.9±1.0**	**91.9±1.0**	**3.13±0.83**
	VLBC- (ours)	82.73±0.14	82.73±0.14	4.67±0.45
	VLBC- \f (ours)	82.7±0.11	82.69±0.11	3.93±0.54
Skin Color				
Task	Method	Accuracy ↑	f1-score ↑	Acc Diff
Gender	Baseline	83.57±0.48	83.56±0.48	2.6±1.1
	Baseline-sampling	84.53±0.13	84.53±0.13	3.4±0.24
	Weighting [38]	83.3±0.28	83.3±0.28	4.2±0.75
	LfF [22]	45.47±0.82	44.21±1.21	0.0±1.07
	CGL-FairHSIC [14]	**91.28±0.26**	**91.28±0.26**	3.17±1.09
	VLBC- (ours)	83.08±0.18	83.06±0.18	**0.03±0.31**
	VLBC- \f (ours)	83.0±0.25	82.98±0.24	0.47±0.56
Age	Baseline	76.25±0.54	76.05±0.57	-12.1±1.06
	Baseline-sampling	76.82±0.51	76.65±0.53	-12.83±0.12
	Weighting [38]	76.17±0.09	76.03±0.12	**-9.93±0.09**
	LfF [22]	32.52±0.73	30.81±0.6	10.17±1.76
	CGL-FairHSIC [14]	**80.55±1.67**	**80.32±1.86**	-11.23±0.4
	VLBC- (ours)	76.52±0.06	76.31±0.05	-11.7±0.45
	VLBC- \f (ours)	76.35±0.43	76.15±0.44	-11.97±0.46

Table 4. Results on the UTKFace dataset (*age* and *skin color*).

We show the results in Tab. 2 and 3, where we compare the baselines with the following SOTA works: Wang et al. [38] (weighting), Learning from Failure [22] (LfF) and Fairness with the Partially annotated Group labels [14] (CGL-FairHSIC). CGL-FairHSIC proposes to improve fairness by incorporating a partially annotated dataset, thus we apply it to the synthetic dataset. Please note that we denote with a "-" the (Δ_A) and (Δ_M) metrics when the model collapses (e.g, f1-score of LfF [22] and CGL-FairHSIC [14] in some cases). The results show how the proposed framework (VLBC-) is always capable of mitigating bias with respect to the baseline model on all attributes and metrics exhibiting *consistency* over multiple settings. The comparison with SOTA methods highlights how other works are robust in some settings but fail in others. Specifically, Wang et al. [38] (weighting) deteriorates the model's fairness in *wearing necktie*, *arched eyebrows*, and *chubby* attributes when *age* is the protected characteristic. LfF [22] is performing poorly, achieving similar or worse bias than the baseline model trained only on the original data and showing a clear drop in accuracy and f1-score with *skin color* as the protected characteristic. CGL-FairHSIC [14] is performing well when *skin color* is the protected characteristic, but fails on *age*. We argue that weighting [38] fails since it possibly uses for training multiple instances of the same image from the minority class. LfF [22] learns the de-biased model by presenting hard samples coming from a second model specifically trained for increasing the bias potentially challenging the model leading to worse performance (see Table 3). Finally, CGL-FairHSIC [14] fails because the fairness loss makes the model converge towards negative predictions degrading in this way the performance.

Moreover, as discussed in Sect. 3, the "baseline-sampling" approach mitigates the bias only when enough images are available for balancing the training set – this is clear in the case of *age* (see Tab. 2). In this setting, the diffusion model has a low bias against *age*, thus it generates enough images for both young and old people to balance the original training set. By contrast, the bias mitigation performance worsens when testing the same approach on *skin color* (see Tab. 3), since not enough images of black people are generated by the generative model. These empirical results confirm our intuition that augmentation is required when samples from the minority class are not enough.

In terms of overall performance, our method is able to mitigate the bias while preserving or improving the accuracy and f1-score of the model, demonstrating its effectiveness. In Tab. 4, we show the results of bias mitigation applied on the UTKFace [42] dataset with *age* and *skin color* as protected characteristics. We note that our method is *consistent* also in this dataset mitigating the bias while preserving the performance. Weighting [38] increases the bias on two attributes out of three and LfF [22] does not preserve the overall performance. CGL-FairHSIC [14] worsens the bias performance on *gender* as attribute and *skin color* as protected characteristic, while showing satisfactory performance on the other two settings. Finally, the "baseline-sampling" approach maintains the existing bias on *age*, while worsening it on *skin color* similarly to what was discussed above. As a final remark, we note that filtering out the images with incorrect protected characteristic augmentation has a low impact on the results (VLBC- \f). This is due to a low error rate during the augmentations.

4.2. Increasing the bias with VLBC+

Remark: *The experiments done in this section are for scientific purposes only and we discourage increasing the bias against specific ethical groups or sensitive attributes.*

As discussed in previous sections, the proposed framework is also capable of increasing the bias towards a specific attribute/protected characteristic, due to its versatile augmentation module (ContraCLIP+DiffAE). We denote this variant of our method as VLBC+. A useful scenario might be that of increasing the downstream task accuracy on particular attributes or the augmentation of a given dataset by generating faces with specific attributes. However, we stress that increasing bias towards specific attributes must be carefully considered and justified to avoid discriminative and unfair practices. For the sake of coherence, we report here the results of this setting applied to CelebA-HQ [21] on the same attributes and protected characteristics discussed above (Sect. 4.1). In this scenario, we augment the images to unbalance, even more, the majority class (e.g., white people). We aim at doubling the majority class; that is, to give enough statistical evidence to the model during training. As we can see in Tab. 5, our method (VLBC+) is able to increase the bias on four attributes out of nine. We argue that, similarly to the *sampling-baseline* approach, VLBC+ fails when not enough images are augmented, thus not doubling the majority class. To further investigate this issue, we report, in Fig. 4, the number of augmented samples compared to the original training set on the attributes where this method is struggling. Since we are augmenting the majority class (e.g., "white" skin colour), we report the number of samples conditioned on the majority protected characteristic class showing the positive or negative number of samples for the downstream classification attribute. For example, the leftmost chart shows the number of samples having (positive) or not having (negative) "straight hair" appearing as "white people" (majority class). As we can see, we cannot generate enough images to actually double the class, since the synthetic dataset $\mathcal{X}_s$ is itself biased and does not capture rare combinations (*attribute-protected characteristics*) to augment. This is due to the inherent bias of the generative model – that is, we may not be able to generate enough black people to augment towards white in order to influence the bias. When this does not happen, our method does increase the bias, as expected in this setting.

Age

Task	Method	Accuracy ↑	Acc Diff
Wearing Necktie	Baseline	94.37±0.11	10.65±0.31
	VLBC+ (ours)	94.35±0.06	**11.15±0.01**
Chubby	Baseline	93.24±0.04	15.85±0.65
	VLBC+ (ours)	93.01±0.05	**17.58±0.09**
Arched Eyebrows	Baseline	80.15±0.3	**-9.0±0.2**
	VLBC+ (ours)	80.0±0.11	-8.8±0.18
Double Chin	Baseline	94.17±0.13	**15.15±0.6**
	VLBC+ (ours)	94.44±0.05	13.73±0.13

Skin Color

Task	Method	Accuracy ↑	Acc Diff
Big Lips	Baseline	65.62±0.54	6.32±1.47
	VLBC+ (ours)	65.13±0.05	**11.66±0.52**
Big Nose	Baseline	78.82±0.07	3.53±0.35
	VLBC+ (ours)	78.68±0.14	**6.57±0.42**
Straight Hair	Baseline	82.52±0.53	**10.17±0.81**
	VLBC+ (ours)	82.92±0.01	9.63±0.01
Young	Baseline	85.25±0.09	**-9.04±0.09**
	VLBC+ (ours)	85.52±0.09	-7.59±0.38
Wearing Necktie	Baseline	94.35±0.11	**-8.1±0.63**
	VLBC+ (ours)	94.3±0.03	-6.98±0.34

Table 5. Results of VLBC employed to increase bias (VLBC+).

4.3. Ablation study

We present our ablation on the number of samples of the minority-protected characteristic to show how the bias can be further controlled. We sample 20%, 40%, 60%, 80%, 100% (whole training set) of the minority class and, finally,

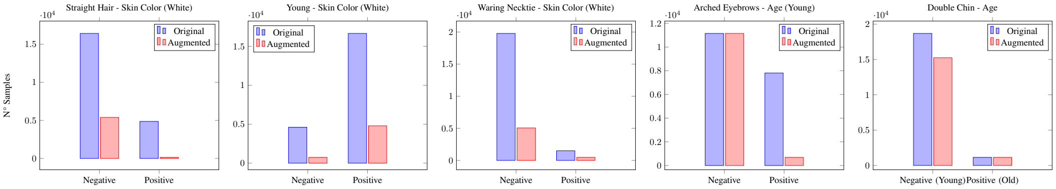
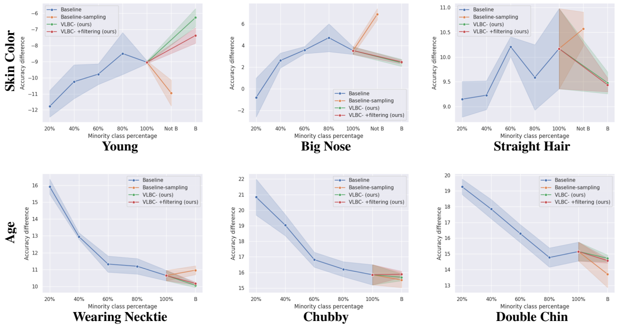


Figure 4. Distribution of the number of samples for the failure cases of Tab. 5.



Young

Big Nose

Straight Hair

Wearing Necktie

Chubby

Double Chin

Figure 5. Accuracy difference for three downstream attributes with *skin color* and *age* as protected characteristics. *B* and *Not B* denote balanced or not balanced data, respectively. Clearly, sampling fails when not enough images are available for sampling.

we plug our augmented data to go beyond the 100% balancing the dataset (as in Sect. 4.1). Fig. 5 shows this ablation on *Baseline*, *Baseline-sampling*, and *VLBC-* with and without *filtering*, considering three downstream tasks with *skin color* and *age* as the protected characteristics. Fig. 5 shows a trend of the difference in accuracy proving that adding data to the training set controls the behavior of the model with respect to the protected characteristic. Moreover, this experiment allows us to study whether our method is able to invert the original trend (increasing bias). The graphs report how our method is consistent in mitigating the bias while the sampling method fails on *skin color* due to the balancing issue discussed earlier (please note that on "Young" the difference in accuracy is negative, thus we want it to increase towards 0). We note that, occasionally, training in low data regime (e.g., 20%) leads to better fairness (e.g., *straight hair*, *big nose*), but only by sacrificing the accuracy.

4.4. Qualitative results

Given a pre-trained generative path, our augmentation module can control the manipulation strength by means of traversal length. In Fig. 3, we show the traversals for the two studied protected characteristics (*skin color* and *age*). Clearly, the longer the traversal length, the stronger the manipulation. This demonstrates the effectiveness of the proposed augmentation module (ContraCLIP+DiffAE) in manipulating facial images toward the desired protected characteristics, while at the same time the overall image quality and other facial attributes are preserved.

5. Limitations

Our work exhibits potential limitations due to the assumptions that: (i) the learnt latent paths convey the desired manipulation while preserving the downstream attribute (disentanglement), and (ii) a good pseudo-labelling module is employed. For (i), we attempt to impose the orthogonality of the paths by employing a contrastive loss which improves their disentanglement, while for (ii) we experimentally show (Sect. 4) that accuracy remains stable across different settings, suggesting that the proposed framework exhibits robustness even with a simple pseudo-labelling module (Sect. 3). Finally, our method requires a generator with an editable space, pre-trained on data where the attributes to be manipulated are well-represented.

6. Conclusion

In this paper, we presented a novel framework for controlling the bias in facial datasets leveraging a *pre-trained* and fixed diffusion model. We built on ContraCLIP [36] in order to find meaningful natural language driven generative paths in the semantic space of DiffAE [27], which we then applied to augment a given dataset with respect to under-represented protected characteristics (e.g., black people), making it fairer for downstream tasks. The proposed bias mitigation method (VLBC-) is able to counteract the bias learnt from a downstream model, while preserving accuracy and showing competitive results against existing SOTA works [14, 22, 38]. Additionally, VLBC- exhibits consistency across multiple settings, a trait missing in concurrent works [14, 22]. Finally, we showed that the proposed framework, besides mitigating bias, is also capable of increasing it (VLBC+), providing full control over bias towards specific attributes. As an interesting future direction, we consider the extension of our method beyond binary classification downstream tasks.

Acknowledgments: This work was supported by the MUR PNRR project FAIR (PE00000013) funded by the NextGenerationEU, the EU project ELIAS (No. 101120237), and the Junior Fellows Program of EU H2020 project AI4Media (No. 951911).

References

[1] Mahed Abroshan, Mohammad Mahdi Khalili, and Andrew Elliott. Counterfactual fairness in synthetic data generation. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022. 2

[2] Sharat Agarwal, Sumanyu Muku, Saket Anand, and Chetan Arora. Does data repair lead to fair models? curating contextually fair data to reduce model bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3298–3307, January 2022. 1

[3] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016. 1

[5] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. HyperReenact: One-shot reenactment via jointly learning to refine and retarget faces. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[6] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Stylemask: Disentangling the style space of stylegan2 for neural face reenactment. In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*, pages 1–8. IEEE, 2023. 2

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. 1

[8] Bhushan Chaudhari, Himanshu Chaudhary, Aakash Agarwal, Kamna Meena, and Tanmoy Bhowmik. Fairgen: Fair synthetic data generation, 2022. 2

[9] Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 915–924, January 2022. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Moreno D'incà, Cigdem Beyan, Radoslaw Niewiadomski, Simone Barattin, and Nicu Sebe. Unleashing the transferability power of unsupervised pre-training for emotion recognition in masked and unmasked facial images. *IEEE Access*, 2023. 1

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2

[13] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[14] Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10348–10357, 2022. 1, 3, 5, 6, 7, 8

[15] Kimmo Karkkainen and Jungseock Joo. 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*, pages 1548–1558, 2021. 3, 4

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 5

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. IEEE, 2020. 4

[18] Runze Li, Tomaso Fontanini, Andrea Prati, and Bir Bhanu. Face synthesis with a focus on facial attributes translation using attention mechanisms. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):76–90, 2023. 1

[19] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 6

[20] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 3, 4, 6, 7

[22] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 5, 6, 7, 8

[23] James Oldfield, Markos Georgopoulos, Yannis Panagakis, Mihalis A Nicolaou, and Ioannis Patras. Tensor component analysis for interpreting the latent space of gans. *arXiv preprint arXiv:2111.11736*, 2021. 2

[24] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, and Ioannis Patras. PandA: Unsupervised learning of parts and appearances in the feature maps of GANs. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[25] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis A. Nicolaou, and Ioannis Patras. Parts of speech-grounded subspaces in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[26] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence. Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, 55(6):1–29, dec 2022. 1

[27] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5, 8

[28] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[31] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. 6

[32] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. Intra-processing methods for debiasing neural networks. In *Advances in Neural Information Processing Systems*, 2020. 1, 2

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[34] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[35] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. 1

[36] Christos Tzelepis, Oldfield James, Georgios Tzimiropoulos, and Ioannis Patras. ContraCLIP: Interpretable GAN generation driven by pairs of contrasting sentences, 2022. 1, 3, 4, 5, 8

[37] Christos Tzelepis, Georgios Tzimiropoulos, and I. Patras. WarpedGANSpace: Finding non-linear RBF paths in GAN latent space. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6373–6382, 2021. 2

[38] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 1, 2, 5, 6, 7, 8

[39] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks, 2018. 2, 3, 6

[40] Weijie Xu, Jinjin Zhao, Francis Iannacci, and Bo Wang. Ffpdg: Fast, fair and private data generation, 2023. 2

[41] Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[42] Zhifei Zhang, Yan Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 2, 6, 7

[43] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 1