Distributionally Generative Augmentation for Fair Facial Attribute Classification

Fengda Zhang[1]*, Qianpei He[1]*, Kun Kuang[1]†, Jiashuo Liu[2],
Long Chen[3], Chao Wu[1], Jun Xiao[1], Hanwang Zhang[4,5]
[1]Zhejiang University [2]Tsinghua University [3]HKUST [4]NTU [5]Skywork AI
{fdzhang,hqp,kunkuang,chao.wu,junx}@zju.edu.cn liujiashuo77@gmail.com
longchen@ust.hk hanwangzhang@ntu.edu.sg

Abstract

Facial Attribute Classification (FAC) holds substantial promise in widespread applications. However, FAC models trained by traditional methodologies can be unfair by exhibiting accuracy inconsistencies across varied data subpopulations. This unfairness is largely attributed to bias in data, where some spurious attributes (e.g., Male) statistically correlate with the target attribute (e.g., Smiling). Most of existing fairness-aware methods rely on the labels of spurious attributes, which may be unavailable in practice. This work proposes a novel, generation-based two-stage framework to train a fair FAC model on biased data without additional annotation. Initially, we identify the potential spurious attributes based on generative models. Notably, it enhances interpretability by explicitly showing the spurious attributes in image space. Following this, for each image, we first edit the spurious attributes with a random degree sampled from a uniform distribution, while keeping target attribute unchanged. Then we train a fair FAC model by fostering model invariance to these augmentation. Extensive experiments on three common datasets demonstrate the effectiveness of our method in promoting fairness in FAC without compromising accuracy. Codes are in https://github.com/heqianpei/DiGA.

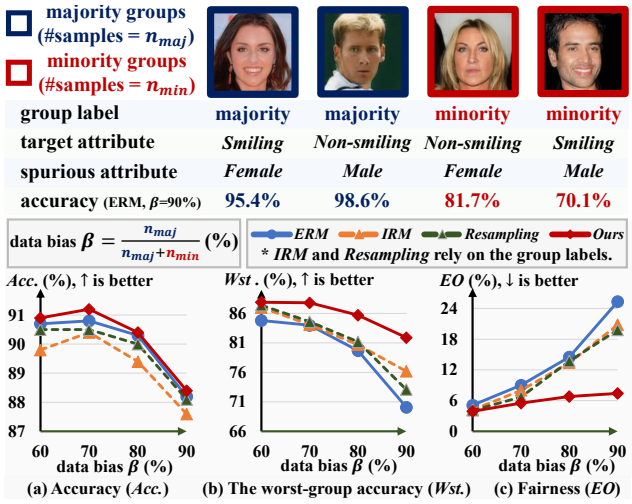

Figure 1. FAC models can be unfair by exhibiting accuracy inconsistencies across varied data subpopulations (*e.g.*, 95.4% accuracy on *Smiling*&*Female* and 70.1% accuracy on *Smiling*&*Male*). This unfairness is predominantly attributed to data bias, measured by β. In general, the more biased the data (*i.e.*, larger β), the more unfair the model. Most of existing methods such as IRM [1] and resampling [59] rely on the labels of spurious attributes. Our method can improve the fairness, measured by *EO* and the worst-group accuracy (Eq. (4) and (5)), of FAC models without additional annotations. Experiments above are performed on CelebA [46].

1. Introduction

Facial Attribute Classification (FAC) has garnered significant interest owing to its broad and practical applications like face verification and image retrieval [66, 88]. The goal of FAC is to predict a certain *target attribute* (*e.g.*, *Smiling*) of a given facial image. Unfortunately, previous studies have shown that the FAC models can be unfair by exhibiting accuracy inconsistencies across different data subpopulations [55]. This unfairness is predominantly attributed to *bias* in the training data [13, 19, 51, 71]. For example, as shown in Figure 1, the majority of *Smiling* images in the training dataset are *Female* (termed as *spurious attribute*). Then, the FAC models trained by traditional methods (*e.g.*, Empirical Risk Minimization (ERM)) may use the spurious attribute as a shortcut to predict the target attribute. As a results, the models may suffer from low accuracy on certain data subpopulations (*e.g.*, *Non-smiling*&*Female*), which seriously hinders their applications in the real world [41].

To train a fair model on the biased dataset, a number of approaches have been proposed. These methods can be broadly divided into two categories. The first category miti-

*These authors contributed equally to this work.
†Corresponding author.

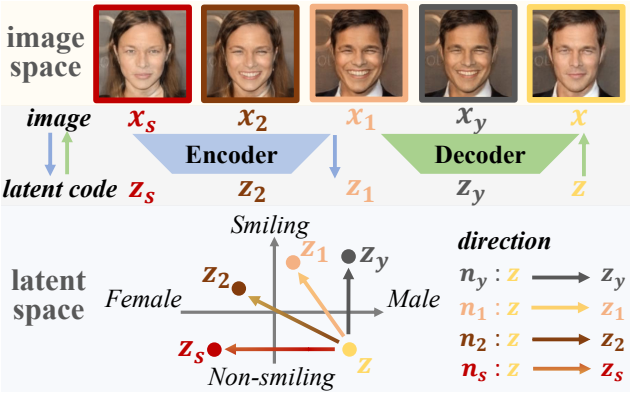

Figure 2. Moving the *latent codes* $\boldsymbol{z}$ of a well-trained generative model in a *learned direction* $\boldsymbol{n}$ can edit the target attribute (*e.g.*, *Smiling*) of images ($\boldsymbol{x} \rightarrow \boldsymbol{x}_y$) [64]. We observe that if the data is biased, the learned direction will be *biased* by containing information of spurious attributes (*e.g.*, *Male*) ($\boldsymbol{x} \rightarrow \boldsymbol{x}_1, \boldsymbol{x}_2$). Based on this, we synthesize a direction $\boldsymbol{n}_s$ to manipulate the spurious attributes while keeping target attribute unchanged ($\boldsymbol{x} \rightarrow \boldsymbol{x}_s$).

gates bias by adding a fairness-aware regularization into the training optimization objective [12, 31, 61]. However, recent findings indicate that the regularization terms of these methods easily suffer from overfitting, causing such methods to degenerate into ERM [26, 86]. The second category transforms the data by reweighting or augmentation to mitigate bias [34, 62, 85]. Among them, a series of recent studies have achieved great success at fair FAC by using generative models to construct an unbiased dataset [39, 56, 58, 81].

Although existing methods have made strides in improving fairness, most of them require the labels of spurious attributes. Unfortunately, spurious attributes annotation may be unavailable in practice for some reasons [2, 6, 22, 37]. Firstly, the vast spectrum of attributes present in images creates a challenge in identifying which are spurious attributes. Secondly, some attributes (*e.g.*, *Attractive*) are difficult to label due to their inherent subjectivity and ambiguity. Lastly, the annotation costs are expensive, especially for large-scale datasets. These considerations naturally prompt the inquiry: ***How can one construct a fair FAC model on biased data without the labels of spurious attributes?***

In this paper, we solve this problem based on a finding in the generative models, as show in Figure 2. Previous studies have shown that *latent codes* for well-trained generative models actually encode disentangled representations, and moving the latent codes $\boldsymbol{z}$ in a *learnable direction* $\boldsymbol{n}$ can manipulate the target attribute (*e.g.*, *Smiling*) of images ($\boldsymbol{x} \rightarrow \boldsymbol{x}_y$) [64, 69]. Based on this, we further observe that the learned direction will be *biased* if there is a spurious correlation in training data. By 'biased', we mean that the direction also contains semantics of spurious attributes (*e.g.*, *Male*), so that the potential spurious attributes will change along with the target attribute during editing ($\boldsymbol{x} \rightarrow \boldsymbol{x}_1, \boldsymbol{x}_2$).

Inspired by this finding, we introduce a two-stage framework to address the posed question. (1) In the first stage, we identify the spurious attributes via generative models. Specifically, we combine two different biased directions (*i.e.*, $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ in Figure 2) in a proper way to cancel out the semantic of target attribute, so that the combined direction $\boldsymbol{n}_s$ will only encode spurious attributes. Then, by editing one or more images with the combined direction, the changes of spurious attributes will be faithfully reflected in the image space, as shown in Figure 2 ($\boldsymbol{x} \rightarrow \boldsymbol{x}_s$). (2) In the second stage, we learn a fair FAC model via generative augmentation. For each image, we first edit its spurious attributes with a random degree sampled from a *uniform distribution*. Then we train a fair FAC model by promoting its invariance to such augmentations. We call the proposed approach ***Distributionally Generative Augmentation*** (***DiGA***).

Our method presents two primary merits. Firstly, no annotation outside of target attribute is required. Leveraging generative models, our method mirrors potential data biases within the image space explicitly, concurrently enhancing interpretability. Secondly, the random degree for fairness-aware generative augmentation follows *uniform distribution*. Compared to the existing *single point* augmentations (*e.g.*, flipping the spurious attributes) [58, 81], it provides more information for the subsequent fair representation learning and thus enhances the representation quality.

We carried out experiments on CelebA [46] and UTKFace [82] datasets for FAC. The classification results in terms of accuracy and fairness show the effectiveness of the proposed *DiGA*. Additionally, we performed extensive analysis experiments to further illustrate the merits of our method in many ways. Moreover, through empirical studies on the Dogs and Cats dataset [33], we showcased the potential of our approach in general bias mitigation.

Our main contributions can be summaried as: (1) The formulation of an interpretable bias detection technique using generative methods for FAC. (2) The introduction of a fair representation learning strategy predicated on distributionally generative augmentation. (3) Comprehensive experiments across three prevalent datasets, demonstrating that our framework effectively enhances fairness without sacrificing accuracy relative to compared baselines in FAC.

2. Related work

Bias Mitigation with Group Information. There are two main branches to train a fair model on biased data. The first branch introduces regularizations into optimization objective [12, 20, 31, 38, 49, 61, 75]. For example, distributionally robust optimization (DRO) methods optimize the worst-case performance [15], while invariant risk minimization (IRM) learns unbiased representations with invariance to different environments [1]. However, regularization-based methods have proved to be prone to the overpes-

simism or overfitting problem [85, 86]. The second branch is to construct an unbiased dataset by transforming data [62, 77]. Typically, reweighting-based methods reweight the data distribution by some heuristics and train models on the reweighted distribution [34, 47, 48]. However, the reweighted distribution still has the same support with the original biased distribution. In order to better improves fairness, recent studies have successfully transformed the training distribution by using generative models to generate training samples for minority groups [39, 56, 58, 81].

Bias Mitigation without Group Information. Some studies have explored how to mitigate bias without additional annotation [2, 3, 6, 22, 43, 44, 60, 76, 84, 87]. Most of these methods predict the bias information as the proxies for the spurious attributes by some heuristics (usually the prediction errors given by a biased classifier) [8, 37, 42, 52]. Recently, researchers have tried to improve robustness based on pre-trained models (e.g., CLIP [57]) without additional annotations [17, 28, 35, 65, 73, 78]. Note that these efforts rely on pre-trained models, and our work can be done without using pre-trained models. For example, we can use a reference model trained by JTT [42] instead of CLIP.

Generative Modeling for Fairness. Generative models have achieved great success in recent years [11, 25, 45, 64, 69]. Some works have proposed to evaluate fairness by generating counterfactual samples [9, 10, 29]. Recently, generation-based methods have demonstrated significant strides in bias mitigation by constructing a balanced and unbiased dataset [27, 39, 56, 58, 81]. However, these methods need the prior of additional annotations. In this paper, we extend the generation-based approach to cases without additional annotations.

Fair Representation Learning. Learning representations is important for reliable performance in visual recognition. Recent years, contrastive learning has been remarkably successful in learning effective representations [4, 5, 18, 24, 53, 70, 74, 80]. However, traditional representation learning methods ignore potential fairness issues. To this end, as a pre-processing method, fair representation learning has achieved great success [7, 14, 50, 54, 63, 71, 79, 83]. For example, ***FSCL*** proposes to learn fair representations by closing the distance of samples with the same target attribute labels but different sensitive attribute labels [55]. However, most of existing fair representation learning methods rely on labels of spurious attributes, while our method avoids this limitation by the proposed bias detection method.

3. Method

In this section, we introduce our two-stage framework to train a fair FAC model on biased data without the labels of spurious attributes. We first state our findings in generative modeling on biased data. Then we propose a generation-based approach for bias detection with theoretical justification. Finally, we develop a method based on distributionally generative augmentation for fair representation learning.

3.1. Findings in Biased Generative Modeling

Image Attribute Manipulation via Latent Space. Previous works have shown that we can manipulate an image's target attribute (*e.g.*, *Smiling*) via latent space of generative models [64, 69]. Typically, given a well-trained GAN model, the generator $G : \mathcal{Z} \to \mathcal{X}$ can map a latent code $\boldsymbol{z} \in \mathcal{Z}$ to an image $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{Z}$ denotes the latent space. As shown in Figure 3(a), we can train a linear classifier in latent space to learn the boundary hyperplane, with a unit normal vector $\boldsymbol{n}$, of the target attribute. Then, the target attribute of image $\boldsymbol{x}$ can be manipulated by altering its latent code $\boldsymbol{z}$ along the normal vector $\boldsymbol{n}$, *i.e.*, $\boldsymbol{x}_{edit} = G(\boldsymbol{z} \pm \alpha \boldsymbol{n})$, where $\alpha \in \mathbb{R}^+$ controls the degree of image attribute editing.

Biased Semantic Direction. Consider that some sensitive attributes (*e.g.*, *Male*) have a statistically association with the target attribute in the training data, as shown in Figure 3(b). In this case, if we train a linear classifier of target attribute in latent space by regularized logistic regression, the learned classifier (*i.e.*, boundary hyperplane) also will be biased. Therefore, when the latent code of the image is moved along the normal of the biased classification hyperplane, not only the target attribute but also the spurious attributes are changed. Moreover, we also note that the biased degree can be affected by the regularization strength.

3.2. Bias Detection via Generative Modeling

Semantic Direction Combination. In order to detect the potential spurious attributes, our idea is to synthesize a direction of spurious attributes in latent space, so that we can manipulate the spurious attributes while keeping the target attribute unchanged in image space. To achieve this, we first train a generative model on the training dataset (or a subset, for efficiency). Then, by using different regularization strengths, we can obtain two different biased semantic directions of target attribute. Finally, as shown in Figure 3(c), we combine these two biased directions by some appropriate combination coefficients. By 'appropriate', we mean that the semantic of the target attribute can be cancelled out to zero while only the semantics of spurious attributes are remained. Theoretical guarantees for the existence of optimal combination coefficients are stated later (Theorem 1). Note that the proposed method naturally supports multi-spurious attribute setting. To extend the method to the multi-class setting, we can transform the problem into multiple binary classification in bias detection stage.

Grid Search for Optimal Combination Coefficients. The remaining question is how to find the optimal combination coefficients. In this paper, we use the grid search method to traverse the parameter values and then apply the different resultant directions to a small image subset of the training

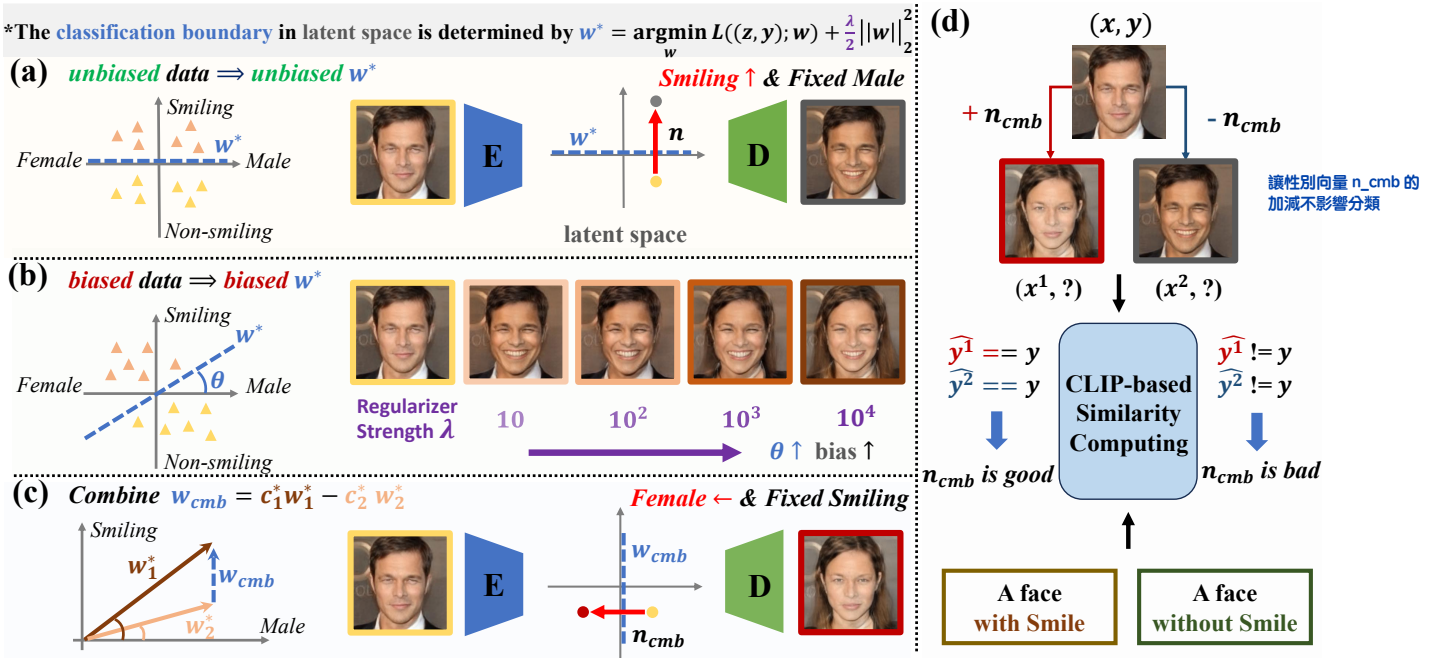

Figure 3. **Illustration of our bias detection method.** (a) Latent codes for well-trained generative models encode disentangled representations, and moving the latent codes along the normal vector $\boldsymbol{n}$ of the learned classification boundary $\boldsymbol{w}^*$ can edit the target attribute of images. (b) The learned boundary $\boldsymbol{w}^*$ will be biased if the training data is biased, and the bias degree of boundary $\boldsymbol{w}^*$ is influenced by the regularization strength λ. (c) By choosing the appropriate coefficients (c_1^*, c_2^*), we can combine two biased boundaries $(\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$ into a new boundary $\boldsymbol{w}_{cmb}$ that is only dependent of the spurious attributes. So the direction $\boldsymbol{n}_{cmb}$, the normal vector of $\boldsymbol{w}_{cmb}$, only encode the semantics of spurious attributes. (d) To find the optimal coefficients (c_1^*, c_2^*), we perform grid search with the help of a reference model.

data. In order to judge whether the components of the target attribute semantics in the direction are cancelled out, we can utilize the artificial judgment or a reference model, as shown in Figure 3(d). More implementation details of grid search for coefficients selection are stated in Experiments.

Remark. We would like to emphasize that previous work has theoretically demonstrated that a model cannot be guaranteed to be fair with only target attribute labels [40], suggesting that introducing additional information is necessary. Note that the cost of the additional information we used is very low. For artificial judgment, only several images needs to be manually judged. As for the reference model, it is not necessary to have high accuracy. There are many open source vision-language foundation models available such as CLIP [57] that can be used to evaluate the changes to the target attributes after editing by zero-shot prediction. Actually, the assumption of reference model is commonly used in fairness studies [6] since it is easy to obtain in practice.

3.3. Theoretical Justification for Bias Detection

Here we provide a theoretical justification for the above bias detection method in a common setting [62]. Without loss of generality, we start from the setup that both target attribute $y \in \{1, -1\}$ and spurious attribute $s \in \{1, -1\}$ are binary. Consider that the training dataset of size n is divided into four groups: two majority groups with $s = y$, each containing $n_{maj}/2$ samples, and two minority groups with $s = -y$, each containing $n_{min}/2$ samples. We define the bias degree of data as $\beta = n_{maj}/(n_{maj} + n_{min}) \in [1/2, 1)$. The larger the β, the stronger the correlation between s and y in the training data. We say the data is unbiased if $\beta = 1/2$. Given a well-trained GAN model, each group has its own distribution over latent codes $\boldsymbol{z} = [\boldsymbol{z}_y, \boldsymbol{z}_s] \in \mathbb{R}^{2d}$ consisting of stable features $\boldsymbol{z}_y \in \mathbb{R}^d$ generated from the target attribute y, and spurious features $\boldsymbol{z}_s \in \mathbb{R}^d$ generated from the spurious attribute s:

這只是理論上的表示，並不是真實的實作過程

$$\boldsymbol{z}_y \mid y \sim N(y\mathbf{1}, \sigma_y^2 I_d), \tag{1}$$

$$\boldsymbol{z}_s \mid s \sim N(s\mathbf{1}, \sigma_s^2 I_d). \tag{2}$$

To get the classification boundary hyperplane, we use regularized logistic regression with optimization objective:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{2d}} \mathbb{E}_{(\boldsymbol{z},y)}[\log(1 + \exp(-y\boldsymbol{w}\boldsymbol{z}))] + \frac{\lambda}{2}||\boldsymbol{w}||_2^2, \tag{3}$$

where $\boldsymbol{w} = [\boldsymbol{w}_y, \boldsymbol{w}_s]$ are linear classifier parameters and $\lambda > 0$ controls regularization strength. The parameters of learned classifier are denotes as $\boldsymbol{w}^* = [\boldsymbol{w}_y^*, \boldsymbol{w}_s^*]$, and we define the bias degree of the classifier as $\beta_{clf} = ||\boldsymbol{w}_s^*||/||\boldsymbol{w}_y^*|| \in [0, +\infty)$. The larger β_{clf} is, the more spurious attribute information the learned classifier uses, and thus the greater

the degree of deviation of the classification boundary hyperplane in latent space. The classifier is unbiased if and only if $\beta_{clf} = 0$. Then we have the following theorem:

Theorem 1 (Optimal combination coefficients' existence)**.** *The learned classifier with optimization objective 3 is biased (i.e., $\beta_{clf} > 0$), if the data is biased (i.e., $\beta > 1/2$). Moreover, there exists optimal combination coefficients $(c_1^*, c_2^*) \in \mathbb{R}_+^2$ such that $\boldsymbol{w}_{cmb} := c_1^* \boldsymbol{w}_1^* - c_2^* \boldsymbol{w}_2^* = [0, 1]$ is dependent of s and independent of y, if $\lambda_1 \ll \lambda_2$, where $\boldsymbol{w}_1^*$ and $\boldsymbol{w}_2^*$ are parameters of linear classifiers trained with regularization strengths λ_1 and λ_2, respectively.*

Please refer to Appendix for proof. The optimal combination coefficients yield a classifier of potential spurious attributes in latent space, independent of the target attribute. Theorem 1 reveals the existence of optimal combination coefficients, laying a foundation for the proposed traversal search-based method for bias detection.

3.4. Bias Mitigation via Generative Augmentation

To prevent the model from learning or amplifying potential bias in the training data, we first learn fair representations that contain as little spurious attributes information as possible. Note that the obtained optimal combined direction $\boldsymbol{n}_{cmb}$ can be used to manipulate the spurious attributes of images while keeping the target attribute unchanged. Following this, our idea is to train a representation model $E(\cdot, \boldsymbol{\phi})$ with invariance to changes in spurious attributes.

We implement this idea based on contrastive learning, as shown in Figure 4. Specifically, for each image $\boldsymbol{x}$ in training dataset with latent code $\boldsymbol{z}$, we perform random augmentation, including not only the traditional strategies $T(\cdot)$ such as random clipping, but also the generative augmentation of spurious attributes via direction $\boldsymbol{n}_{cmb}$. By this way, in each iteration, we can get the augmented positive sample pair $\boldsymbol{x}' = T(G(\boldsymbol{z} + \alpha' \boldsymbol{n}_{cmb}))$ and $\boldsymbol{x}'' = T(G(\boldsymbol{z} - \alpha'' \boldsymbol{n}_{cmb}))$, where α' and α'' are uniformly sampled from $[\alpha_l, \alpha_u]$, and $\alpha_l, \alpha_u \in \mathbb{R}^+$ are hyperparameters controlling the variation range of spurious attributes' semantics. Then, we train a fair encoder $E(\cdot, \boldsymbol{\phi})$ by minimizing the distance between representations of positive samples on training dataset. Finally, we train a linear classifier $C(\cdot, \boldsymbol{\omega})$ on the top of frozen encoder $E(\cdot, \boldsymbol{\phi})$ on training dataset for fair classification.

Notably, The degree for spurious attributes manipulation follows a uniform distribution rather than a single point. This distributionally generative augmentation provides finer supervision information as guidance for fair representation learning, thus helping to improve representation quality. We used techniques such as momentum update and stopping gradient like previous works [5, 18] in our implementation.

性別向量 n_cmb 的加減不影響分類

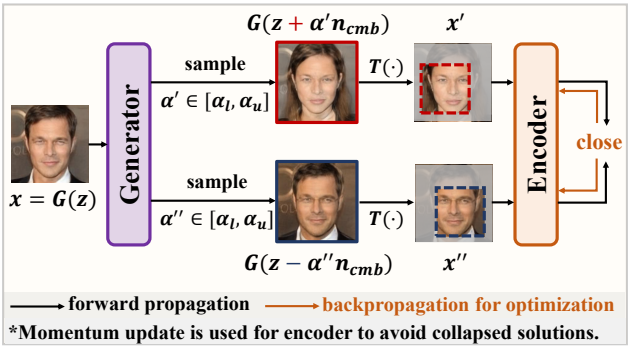

Figure 4. **Distributionally generative augmentation for fair representation learning.** For each image, we edit its spurious attributes by using the combined semantic direction $\boldsymbol{n}_{cmb}$ in latent space. The editing degrees α' and α'' are randomly sampled from a uniform distribution. We also perform traditional augmentation $T(\cdot)$ such as random clipping. The encoder is trained to learn fair and effective representations by closing the distance between augmented views. We use momentum encoder to avoid collapsing.

α 並不是直接選一個值，而是從 uniform distr 裡面來 sample

4. Experiments

4.1. Datasets

To validate our method, we did experiments on identical three datasets that were utilized in prior studies [55, 58, 81].

CelebA [46] is a common dataset used for FAC. Each image has 40 binary attributes labels. Following the setting of the previous works [42, 58, 72], we select *Smiling*, *Blond Hair*, *Black Hair*, *Male* and *Young* as the target attributes, and set *Male* and *Young* as spurious attributes. Besides, to verify the performance of our method in the setting of multi-target labels and multi-spurious attributes, we also set {*Blond Hair*, *Black Hair*} as target attributes and {*Male*, *Young*} as spurious attributes respectively. For each experiment, we randomly sample a biased subset as training dataset with size of 20,000 images, where the majority group and minority group have 90% and 10% of the sample size respectively. We report performance on the whole original test dataset.

UTK-Face [82] contains over 20k facial images, each with attributes labels. Following the experimental setup in the previous works [30, 55], we define a binary spurious attribute *Ethnicity* based on whether the facial image is white or not. The task is to predict the *Gender*. We randomly sample a biased subset consisting of 10,000 images, with the same bias degree as CelebA. We also construct a balanced and unbiased test dataset consisting of 3,200 images.

Dogs and Cats dataset is widely used for general bias mitigation [33, 55, 81]. It contains dog and cat images with additional annotations for partial images about the color of dog/cat is bright or dark. The task is to predict if the image is a cat or a dog and the spurious attribute is color. The biased training set consists of 400 bright cat images, 3,600 bright dog images, 3,600 dark cat images, and 400 dark cat

Figure 5. **Bias detection results on CelebA dataset.** The constructed training dataset is biased, where 75% *Smiling* images are *Female&Young* and 75% *Non-smiling* images are *Male&Non-young*. We only have labels of target attribute *Smiling*. By utilizing the proposed bias detection method, we obtain the combined direction and edit training images. It can be observed that the changes of gender and age are faithfully reflected in image space, illustrating which attributes are spurious attributes explicitly and thus enhancing interpretability.

images. The balanced test set consists of 2,400 images.

4.2. Evaluation Metrics

Our goal is to learn a fair and accurate FAC model. In this paper, we use equalized odds (*EO*) [21], one of the most commonly used notion of group fairness [16], as the fairness metric. Following [30], we extend *EO* to multi-target attribute and multi-spurious attribute setting:

改變 spurious，預測的 probabilistic 應該依舊不會改變。

$$\max_{\substack{\forall y,\hat{y}\in\mathcal{Y}\\ \forall s^i,s^j\in\mathcal{S}}} \left|P_{s^i}(\hat{Y}=\hat{y}\mid Y=y)-P_{s^j}(\hat{Y}=\hat{y}\mid Y=y)\right|, \quad (4)$$

where Y is ground truth, $\hat{Y}$ is predictive label given by the classifier, and $s^i, s^j \in \mathcal{S}$ is the value of spurious attributes. A smaller *EO* means a fairer classifier. We also report the worst-group accuracy defined as:

$$\min_{\substack{\forall y\in\mathcal{Y}\\ \forall s\in\mathcal{S}}} P_s(\hat{Y}=y\mid Y=y), \quad (5)$$

Besides, we use accuracy (%) to measure the model utility.

4.3. Bias Detection Results on Facial Datasets

Bias Detection Results. Consider that the task is to predict whether a given facial image is *Smiling* or not. The training dataset is constructed to be biased, where the target attribute *Smiling* statistically correlates with two potential spurious attributes *Male* and *Young*. Note that only target attribute labels are available during training. We use the proposed bias detection method to obtain the combined semantic direction, and edit the training images to detect the bias. The results on CelebA are shown in Figure 5. We can observe that the changes of the potential spurious attributes *Male* and *Young* are faithfully reflected in the image space.

c_1/c_2	0.5	0.6	0.7	0.8	**0.9**	1.0	1.1	1.2	1.3	1.4	1.5
consistency (%)	73.0	76.0	78.5	83.0	**84.5**	83.5	83.0	82.5	81.0	77.5	76.5

Table 1. Grid search results for optimal combination coefficients.

Implementations. We perform grid search for the optimal combination coefficients c_1^*, c_2^* such that the combined direction $c_1^*\boldsymbol{n}_1 - c_2^*\boldsymbol{n}_2$ only contains semantics of spurious attributes. The criterion is to make the target attribute match most consistent before and after editing, that is, to make the changes of the target attributes as little as possible. In order to improve efficiency, we randomly sample only 100 images from the training dataset. For each search, we edit them positively and negatively respectively by the combined direction to get 200 edited images. Then we predict their target labels by using CLIP as reference model with prompts "A face with/without smile". For ease of search, we rewrite the combined direction as $c_1(\boldsymbol{n}_1 - c_2/c_1\boldsymbol{n}_2)$. So we only need to search for c_2/c_1 (from 0.5 to 1.5 with unit 0.1). The results are shown in Table 1. We set $c_1/c_2 = 0.9$, which makes the target attribute change the least after editing.

4.4. Classification Results on Facial Datasets

Main Results. The classification results on CelebA and UTKFace datasets are shown in Table 2. We measure classification accuracy (*Acc.*), the worst-group accuracy (*Wst.*), and *EO* of trained FAC models. We find that *ERM* achieves good accuracy but suffer from severe unfairness. We also compare our method with various state-of-the-art debiasing baselines that do not require spurious attribute labels including regularization-based methods (*CVaR DRO* [38] and *EIIL* [8]) and reweighting-based methods (*LfF* [52], *JTT* [42], and *MAPLE* [85]). We find that although these debiasing methods improve fairness to some extent, they sacrifice accuracy more or less. Compared with them, our

Method	T=*s* / S=*m*			T=*s* / S=*y*			T=*b* / S=*m*			T=*a* / S=*y*			T=*m* / S=*y*			T=*y* / S=*m*			T=*b*&*a*&*r* / S=*m*			T=*s* / S=*m*&*y*			T=*g* / S=*e*		
	Acc.	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*
ERM [23]	88.2	70.1	25.3	88.3	71.5	15.6	84.2	73.3	17.1	82.8	70.1	19.4	97.2	92.8	5.4	77.7	42.0	52.0	90.6	69.3	24.1	87.3	60.4	33.8	91.4	83.5	12.2
CVaR DRO [38]	87.3	74.0	22.8	87.0	76.1	13.9	84.0	73.9	15.5	81.4	71.8	15.2	96.5	93.0	5.3	75.4	42.3	48.8	90.0	71.8	22.0	86.3	64.0	28.4	90.6	84.5	11.9
EIIL [8]	87.9	75.6	19.7	87.9	72.5	13.3	84.1	73.9	15.7	81.9	73.3	14.4	96.2	93.3	4.9	77.5	45.6	39.2	90.4	71.5	22.0	86.4	60.8	19.7	89.2	84.3	8.3
LfF [52]	87.1	77.5	17.0	85.3	72.9	14.3	84.0	74.0	15.1	82.4	72.5	14.2	97.1	92.9	5.1	77.4	44.2	43.6	89.8	70.8	20.5	85.0	62.5	26.6	86.7	84.6	11.1
JTT [42]	88.0	74.8	19.4	87.6	73.3	14.2	83.9	74.1	16.7	81.1	71.1	16.6	97.0	92.4	5.8	76.3	43.6	47.7	88.3	69.1	23.3	87.3	61.0	31.0	90.5	85.0	10.4
MAPLE [85]	88.1	72.0	19.6	88.1	73.6	13.6	83.7	73.9	14.7	82.4	74.7	13.8	97.1	92.9	4.8	76.3	46.2	43.5	89.9	72.8	18.6	86.0	64.8	31.2	89.4	85.3	9.4
DiGA (ours)	**88.4**	**81.9**	**7.4**	**89.1**	**78.5**	**9.5**	**84.5**	**74.5**	**13.5**	**83.6**	**78.6**	**10.8**	**97.4**	**94.8**	**4.3**	**80.0**	**51.3**	**33.3**	**90.7**	**79.7**	**15.8**	**88.4**	**75.8**	**15.6**	**92.7**	**89.0**	**6.8**

Table 2. **Classification results on facial datasets.** We use classification accuracy (*Acc.*), the worst-group accuracy (*Wst.*), and equalized odds (*EO*) to measure the performance of FAC model on CelebA and UTKFace datasets. T and S represent target and spurious attributes, respectively. *s*, *b*, *a*, *r*, *m*, *y*, *g*, and *e* respectively denote *Smiling*, *Blond Hair*, *Black Hair*, *Brown Hair*, *Male*, *Young*, *Gender*, and *Ethnicity*. The spurious attribute labels are unavailable for all methods during training. All the results are the averaged scores over five runs.

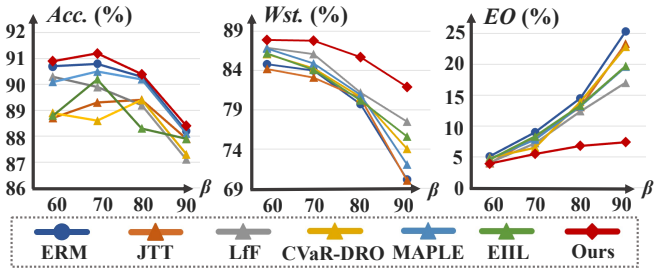

Figure 6. **Classification results on CelebA dataset under different degrees of data bias.** Data bias $\beta := \frac{n_{maj}}{n_{maj}+n_{min}}$, where n_{maj} and n_{min} respectively denote the sample size of majority and minority groups.

Method	label ratio=50%			label ratio=25%			label ratio=10%		
	Acc.	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*
ERM	87.5	67.8	26.1	87.1	65.9	27.7	86.9	62.8	28.9
CVaR DRO	86.6	72.9	22.1	86.6	72.3	22.4	85.5	69.1	27.3
EIIL	86.2	71.3	22.5	85.9	69.6	25.4	86.8	64.2	26.7
LfF	86.9	75.5	19.4	85.9	72.1	23.6	85.5	66.1	27.7
JTT	87.3	72.9	20.1	86.7	71.1	20.6	86.8	67.1	23.1
MAPLE	87.4	73.7	23.8	87.0	72.7	24.2	85.6	69.2	27.1
DiGA (ours)	**88.4**	**81.1**	**7.8**	**88.4**	**78.3**	**8.0**	**88.3**	**78.8**	**8.4**

Table 3. **Classification results on CelebA dataset** (T=*s*, S=*m*) **under settings of incomplete target labels.** We set the label ratio of target attribute as 50%, 25%, and 10% respectively.

DiGA achieves better performance in terms of accuracy, the worst-group accuracy, and *EO* in various settings.

Robustness to Data Bias Degree. In Figure 6, we show the robustness of different algorithms to the degree of data bias β on CelebA (T=*s*, S=*m*), where β is defined as the proportion of the majority group sample to the total sample. We can observe that our method achieves state-of-the-art performance in terms of accuracy and fairness under various data bias degrees. Moreover, as the degree of data bias increases, the *EO* gap between the compared baselines and our method gradually increases. This shows that our method has better robustness to the degree of data bias.

Semi-supervised Classification Results. Our approach relies only on the target labels without the need for additional annotations. However, target labels are not always available due to annotation costs, which motivates us to test the performance of different methods under the setting of incomplete target labels. We set different label ratios, as shown in Table 3. Compared with baselines, our method has better robustness to ratio of target labels. The reason may be that our method trains encoder in self-supervised way, while the target labels are only used to train the linear classifiers.

4.5. T-SNE Visualization

To further explain how our method works, we provide visualization of the learned representations via t-SNE [68] in Figure 7. We divide the CelebA dataset into four groups in terms of target and spurious attributes and randomly sample 500 images from each group. We find that traditional representation learning method BYOL [18] learns information of spurious attributes, so that the representations trained by BYOL can be divided by the spurious attributes. In contrast, the representations learned by ours contain less information of spurious attributes, thus contributing to fair classification.

4.6. Ablation Studies

Ablation Studies on Generative Augmentation. The ablation study results of generative augmentation on CelebA dataset (T=*s*, S=*m*) are shown in Figure 8. Our approach has advantages in the following aspects: (1) *Comparison with traditional augmentation.* Compared with the typical contrastive learning BYOL [18] that only performs traditional augmentation (*e.g.*, random cropping), our method achieves better accuracy and fairness thanks to the fairness-aware generative augmentation. (2) *Comparison with single point augmentation.* Compared with single point generative augmentation [81], our distributionally generative augmentation achieves better accuracy and fairness. Because it considers a wider data distribution and provides more supervision information for fair representation learning. (3) *Trade-off between accuracy and fairness.* We can flexibly balance accuracy and fairness via the range of augmentation degree, while larger degree result in a fairer model.

Ablation Studies on Sampling Ratio for Efficient Model-

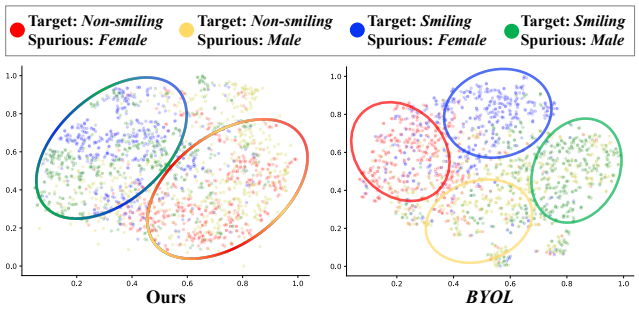

Figure 7. **T-SNE visualization for the learned representations on CelebA** (T=*s*, S=*m*). Compared with BYOL, the representations trained by ours contain less information of spurious attributes.

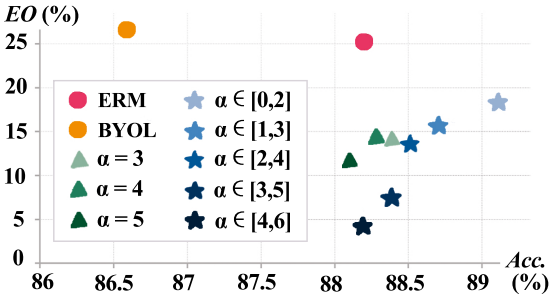

Figure 8. **Ablation studies on generative augmentation.** ▲ and ★ respectively denote single point generative augmentation strategy and our distributionally generative augmentation strategy.

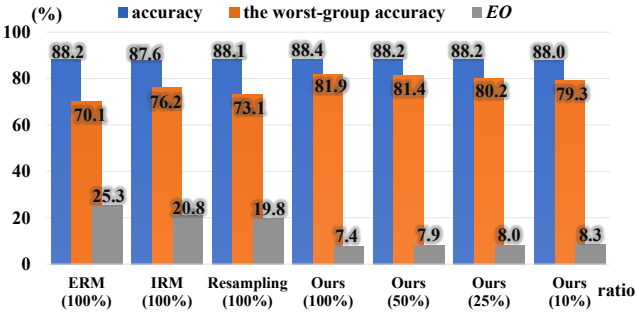

Figure 9. **Empirical studies of sample ratio used for generative modeling.** For efficiency, we can randomly sample a subset to train the generative model, with little final performance degradation.

	Acc.	*Wst.*	*EO*
λ_1=2e-4 λ_2=5e+3	88.3	82.7	9.3
λ_1=1e-4 λ_2=1e+4	88.4	81.9	7.4
λ_1=2e-5 λ_2=5e+4	88.8	85.1	4.8
λ_1=1e-6 λ_2=1e+6	88.8	82.3	7.4

Table 4. **Ablation studies of regularization strength λ_1, λ_2 on CelebA** (T=*s*, S=*m*). We set several groups of λ_1, λ_2, where λ_1 is much smaller than λ_2.

	T=*s* / S=*c*		
Method	*Acc.*	*Wst.*	*EO*
ERM	87.5	67.8	26.1
CVaR DRO	86.6	72.9	22.1
EIIL	86.2	71.3	22.5
LfF	86.9	75.5	19.4
JTT	87.3	72.9	20.1
MAPLE	87.4	73.7	23.8
DiGA (ours)	**88.4**	**81.1**	**7.8**

Table 5. **Classification results on non-facial dataset Dogs and Cats.** *s* denotes the target attribute *Species* and *c* denotes the spurious attribute *Color*.

ing. To improve the efficiency of generative modeling, we implement the sampling strategy. Specifically, we randomly sample a subset to train the generative model and obtain the combined semantic direction. The final classification results on CelebA dataset (T=*s*, S=*m*) are shown in Figure 9. We can observe that the performance of our method decreases very little even when the sampling ratio is small (*e.g.*, 10%).

Ablation Studies on Regularization Strength. In Table 4, we show the classification results by using different regularization strengths. It shows that our method can achieve good results as long as λ_1 is much smaller than λ_2.

4.7. General Bias Mitigation on Non-facial Dataset

To verify the effectiveness of our method on visual data other than faces, we perform experiments on Dogs and Cats dataset, where the target and spurious attributes are *Species* and *Color* respectively. The classification results are shown in Table 5. Our method achieves better accuracy and fairness than other compared baselines, and it bodes well for the potential of our approach for general bias mitigation.

5. Conclusions

In this paper, we proposed a generation-based two-stage framework to train a fair FAC model on biased data without additional annotations. In the first stage, we detect the spurious attributes via generative models. Our method enhances interpretability by explicitly representing the spurious attributes in the image space. In the second stage, for each image, we first edit its spurious attributes, where the editing degree follows a uniform distribution. Then we train a fair FAC model by promoting its invariance to these augmentation. Extensive experiments on the three datasets demonstrate the effectiveness of our approach. In future work, we aim to extend our method to support various visual data, with the help of rapidly developing generative models.

Acknowledgment This work was supported by the National Natural Science Foundation of China (62337001, 62376243, 62037001, U20A20387), the Fundamental Research Funds for the Central Universities (No. 226-2022-00051), the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), Project by Shanghai AI Laboratory (P22KS00111) and National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-022). Long Chen is supported by HKUST Special Support for Young Faculty (F0927) and HKUST Sports Science and Technology Research Grant (SSTRG24EG04).

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2

[2] Carolyn Ashurst and Adrian Weller. Fairness without demographic data: A survey of approaches. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–12. 2023. 2, 3

[3] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3, 5

[6] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pages 1887–1898. PMLR, 2020. 2, 3, 4

[7] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019. 3

[8] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021. 3, 6, 7

[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*, pages 915–924, 2022. 3

[10] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439*, 2019. 3

[11] Yahya Dogan and Hacer Yalim Keles. Semi-supervised image attribute editing using generative adversarial networks. *Neurocomputing*, 401:338–352, 2020. 3

[12] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018. 2

[13] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020. 1

[14] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34:12091–12103, 2021. 3

[15] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018. 2

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 6

[17] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022. 3

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3, 5, 7

[19] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20370–20382, 2023. 1

[20] Tobias Hänel, Nishant Kumar, Dmitrij Schlesinger, Mengze Li, Erdem Ünal, Abouzar Eslami, and Stefan Gumhold. Enhancing fairness of visual attribute predictors. In *Proceedings of the Asian Conference on Computer Vision*, pages 1211–1227, 2022. 2

[21] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 6

[22] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018. 2, 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 3

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[25] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019. 3

[26] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018. 2

[27] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*, 2020. 3

[28] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022. 3

[29] Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd international workshop on fairness, accountability, transparency and ethics in multimedia*, pages 1–5, 2020. 3

[30] 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. 5, 6

[31] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011. 2

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[33] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 2, 5

[34] Dohyung Kim, Sungho Park, Sunhee Hwang, and Hyeran Byun. Fair classification by loss balancing via fairness-aware batch sampling. *Neurocomputing*, 518:231–241, 2023. 2, 3

[35] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Bias-to-text: Debiasing unknown visual biases through language interpretation. *arXiv preprint arXiv:2301.11104*, 2023. 3

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[37] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020. 2, 3

[38] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020. 2, 6, 7

[39] Lei Li, Fan Tang, Juan Cao, Xirong Li Li, and Danding Wang. Bias oriented unbiased data augmentation for cross-bias representation learning. *Multimedia Systems*, 29: 725–738, 2023. 2, 3

[40] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542, 2022. 4

[41] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10052–10061, 2019. 1

[42] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 3, 5, 6, 7

[43] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pages 6804–6814. PMLR, 2021. 3

[44] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Kernelized heterogeneous risk minimization. *arXiv preprint arXiv:2110.12425*, 2021. 3

[45] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019. 3

[46] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 1, 2, 5

[47] Zheqi Lv, Wenqiao Zhang, Shengyu Zhang, Kun Kuang, Feng Wang, Yongwei Wang, Zhengyu Chen, Tao Shen, Hongxia Yang, Beng Chin Ooi, et al. Duet: A tuning-free device-cloud collaborative parameters generation framework for efficient device model generalization. In *Proceedings of the ACM Web Conference 2023*, pages 3077–3085, 2023. 3

[48] Zheqi Lv, Wenqiao Zhang, Zhengyu Chen, Shengyu Zhang, and Kun Kuang. Intelligent model update strategy for sequential recommendation. In *Proceedings of the ACM Web Conference 2024*, 2024. 3

[49] Jiali Ma, Zhongqi Yue, Kagaya Tomoyuki, Suzuki Tomoki, Karlekar Jayashree, Sugiri Pranata, and Hanwang Zhang. Invariant feature regularization for fair face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20861–20870, 2023. 2

[50] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018. 3

[51] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021. 1

[52] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 3, 6, 7

[53] Sungho Park, Dohyung Kim, Sunhee Hwang, and Hyeran Byun. Readme: Representation learning by fairness-aware disentangling method. *arXiv preprint arXiv:2007.03775*, 2020. 3

[54] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-aware information align-

ment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2403–2411, 2021. 3

[55] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022. 1, 3, 5

[56] Momchil Peychev, Anian Ruoss, Mislav Balunović, Maximilian Baader, and Martin Vechev. Latent space smoothing for individually fair representations. In *European Conference on Computer Vision*, pages 535–554. Springer, 2022. 2, 3

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4

[58] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9301–9310, 2021. 2, 3, 5

[59] Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. *Advances in neural information processing systems*, 33:361–371, 2020. 1

[60] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. What's in a name? reducing bias in bios without access to protected attributes. *arXiv preprint arXiv:1904.05233*, 2019. 3

[61] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 2

[62] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020. 2, 3, 4

[63] Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*, 2021. 3

[64] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 3

[65] Joonghyuk Shin, Minguk Kang, and Jaesik Park. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023. 3

[66] Nathan Thom and Emily M Hand. Facial attribute recognition: A survey. *Computer Vision: A Reference Guide*, pages 1–13, 2020. 1

[67] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 3

[68] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[69] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[70] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 3

[71] 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, 3

[72] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2022. 5

[73] Olivia Wiles, Isabela Albuquerque, and Sven Gowal. Discovering bugs in vision models using off-the-shelf image generation and captioning. *arXiv preprint arXiv:2208.08831*, 2022. 3

[74] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020. 3

[75] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 578–586, 2021. 2

[76] Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020. 3

[77] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 813–822, 2022. 3

[78] Jianhao Yuan, Francesco Pinto, Adam Davies, and Philip Torr. Not just pretty pictures: Toward interventional data augmentation using text-to-image generators. *arXiv preprint arXiv:2212.11237*, 2022. 3

[79] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013. 3

[80] Fengda Zhang, Kun Kuang, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Yueting Zhuang, and Xiaolin Li. Federated unsupervised representation learning. *arXiv preprint arXiv:2010.08982*, 2020. 3

[81] 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, 5, 7

[82] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017. 2, 5

[83] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32, 2019. 3

[84] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1433–1442, 2022. 3

[85] Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. Model agnostic sample reweighting for out-of-distribution learning. In *International Conference on Machine Learning*, pages 27203–27221. PMLR, 2022. 2, 3, 6, 7

[86] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022. 2, 3

[87] Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. 2023. 3

[88] Ni Zhuang, Yan Yan, Si Chen, and Hanzi Wang. Multi-task learning of cascaded cnn for facial attribute classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2069–2074. IEEE, 2018. 1

Distributionally Generative Augmentation for Fair Facial Attribute Classification

Supplementary Material

The supplementary materials are organized as follows:

- In Appendix A, we give the proof for Theorem 1. Theorem 1 guarantees the existence of optimal combination coefficients, so that we can use grid search to find them;
- In Appendix B, as an empirical supplement to Theorem 1, we show our observations on synthetic dataset to reveal the relationship between β_{clf} (the bias of learned classifier in latent space) and λ (the regularization strength);
- In Appendix C, we present the additional results of bias detection on real facial dataset to more intuitively show why and how our approach works.
- In Appendix D, we present the implementation details.

Appendix A. Proof for theoretical justification

Proof: We first define the sample ratio of majority group and minority group as $p_{maj} = n_{maj}/(n_{maj} + n_{min})$ and $p_{min} = n_{min}/(n_{maj}+n_{min})$ respectively. The optimization objective $R(\boldsymbol{w})$ can be written as

$$\begin{aligned}
R(\boldsymbol{w}) =& \mathbb{E}_{(\boldsymbol{z},y)}[\log(1+\mathrm{e}^{-\mathrm{y}\boldsymbol{w}\boldsymbol{z}})] + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \\
=& \frac{p_{maj}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{-\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{p_{maj}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\text{-}\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\text{-}\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\text{-}\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{-\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{\boldsymbol{z}_y\sim N(\text{-}\mathbf{1},\sigma_y^2 I_d)}\mathbb{E}_{\boldsymbol{z}_s\sim N(\mathbf{1},\sigma_s^2 I_d)}[\log(1+\mathrm{e}^{\boldsymbol{w}\boldsymbol{z}})] \\
&+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2.
\end{aligned} \tag{6}$$

Without loss of generality, we let $d = 1$. Then we have

$$\begin{aligned}
R(\boldsymbol{w}) =& \frac{p_{maj}}{2}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}-\mathrm{w_s z_s}})] \\
&+\frac{p_{maj}}{2}\mathbb{E}_{z_y\sim N(-1,\sigma_y^2),z_s\sim N(-1,\sigma_s^2)}[\log(1+\mathrm{e}^{\mathrm{w_y z_y}+\mathrm{w_s z_s}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(-1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}-\mathrm{w_s z_s}})] \\
&+\frac{p_{min}}{2}\mathbb{E}_{z_y\sim N(-1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{\mathrm{w_y z_y}+\mathrm{w_s z_s}})] \\
&+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \\
=& p_{maj}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}-\mathrm{w_s z_s}})] \\
&+p_{min}\mathbb{E}_{z_y\sim N(1,\sigma_y^2),z_s\sim N(1,\sigma_s^2)}[\log(1+\mathrm{e}^{-\mathrm{w_y z_y}+\mathrm{w_s z_s}})] \\
&+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2.
\end{aligned} \tag{7}$$

For convenience, we write $\mathbb{E}_{z_y\sim N(1,\sigma_y^2)}$ and $\mathbb{E}_{z_s\sim N(1,\sigma_s^2)}$ as $\mathbb{E}_{z_y}$ and $\mathbb{E}_{z_s}$ respectively without causing any ambiguity. Our goal is to minimize $R(w_y, w_s)$. So we focus on the gradients of classifier parameters w_y and w_s:

$$\begin{aligned}
&\nabla_{w_y} R(w_y, w_s) \\
=& p_{maj}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y + w_s z_s})}(-z_y)] \\
&+p_{min}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y - w_s z_s})}(-z_y)] \\
&+\lambda w_y
\end{aligned} \tag{8}$$

and

$$\begin{aligned}
&\nabla_{w_s} R(w_y, w_s) \\
=& p_{maj}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y + w_s z_s})}(-z_s)] \\
&+p_{min}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y z_y - w_s z_s})}z_s] \\
&+\lambda w_s.
\end{aligned} \tag{9}$$

We use proof by contradiction. Let w_s^* be zero. Then we have

$$\begin{aligned}
&\nabla_{w_s} R(w_y^*, 0) \\
=& p_{maj}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}(-z_s)] \\
&+p_{min}\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
=& (-p_{maj})\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
&+(1-p_{maj})\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
=& (1-2p_{maj})\mathbb{E}_{z_y}\mathbb{E}_{z_s}[\frac{1}{(1+e^{w_y^* z_y})}z_s] \\
=& (1-2p_{maj})\mathbb{E}_{z_y}[\frac{1}{(1+e^{w_y^* z_y})}]\mathbb{E}_{z_s}[z_s] \\
=& (1-2p_{maj})\mathbb{E}_{z_y}[\frac{1}{(1+e^{w_y^* z_y})}] \\
<& 0.
\end{aligned} \tag{10}$$

Note that $\mathbb{E}_{z_y}[\frac{1}{(1+e^{w_y^* z_y})}] > 0$, so that the $\nabla_{w_s} R(w_y^*, 0) = 0$ if and only if the majority group sample ratio $p_{maj} = 1/2$ (*i.e.*, the data is unbiased). The above equation shows that the solution w_s^* cannot be zero. Similarly, we also have

$$\nabla_{w_y} R(0, w_s^*) < 0. \tag{11}$$

So the bias degree of the classifier $\beta_{clf} = \|w_s^*\|/\|w_y^*\| > 0$ if the data is biased (*i.e.*, $\beta = p_{maj} => 1/2$). Different values of λ will scale the impact of the regularization term, affecting the solution $\boldsymbol{w}^\star = (w_y^*, w_s^*)$ of logistic regression. Denote the solutions under regularization strength λ_1 and λ_2 are $\boldsymbol{w_1^\star} = (w_{y1}^*, w_{s1}^*)$ and $\boldsymbol{w_2^\star} = (w_{y2}^*, w_{s2}^*)$ respectively. As we have proven before, w_{y1}^*, w_{s1}^*, w_{y2}^*, and w_{s2}^* are not zero. Then we construct $c_1^* = w_{y2}^*/(w_{y2}^* w_{s1}^* - w_{y1}^* w_{s2}^*)$ and $c_2^* = w_{y1}^*/(w_{y2}^* w_{s1}^* - w_{y1}^* w_{s2}^*)$ such that $\boldsymbol{w}_{cmb} := c_1^* \boldsymbol{w}_1^* - c_2^* \boldsymbol{w}_2^* = [0, 1]$. Here we have completed the proof of the existence of the optimal combination coefficients. □

Appendix B. Observations on synthetic dataset

In this section, as an empirical supplement to Theorem 1, we explore the relationship between β_{clf} (bias of learned linear classifier in the latent space) and λ (regularization strength used in logistic regression) on synthetic dataset.

Experimental Setup. Following the previous studies [62], we use the same settings as in the theoretical justification. Specifically, target attribute $y \in \{1, -1\}$ and spurious attribute $s \in \{1, -1\}$ are binary. The training dataset contains $n = 20000$ samples, which can be divided into four groups: two majority groups with $s = y$, each containing $n_{maj}/2$ samples, and two minority groups with $s = -y$, each containing $n_{min}/2$ samples. In the latent space of generative models, each group has its own distribution over latent codes $\boldsymbol{z} = [\boldsymbol{z}_y, \boldsymbol{z}_s] \in \mathbb{R}^{200}$ consisting of stable features $\boldsymbol{z}_y \in \mathbb{R}^{100}$ generated from the target attribute y, and spurious features $\boldsymbol{z}_s \in \mathbb{R}^{100}$ generated from the spurious attribute s: $\boldsymbol{z}_y \mid y \sim N(y\mathbf{1}, \sigma_y^2 I_{100})$ and $\boldsymbol{z}_s \mid s \sim N(s\mathbf{1}, \sigma_s^2 I_{100})$. To get the classification boundary, we use logistic regression with regularization strength λ. Recall that the bias degree of the classifier as $\beta_{clf} = \|\boldsymbol{w}_s^*\|/\|\boldsymbol{w}_y^*\| \in [0, +\infty)$. We set different data bias by using different ratios $n_{maj} : n_{min}$. We also set different standard deviations for $\boldsymbol{z}_y$ and $\boldsymbol{z}_s$. All results were averaged over 100 random repetitions.

Observations. As shown in Table 6, in most cases, if we increase the regularization strength λ in logistic regression, the classifier bias β_{clf} will be larger. This observation motivates us to design a ***simple*** but ***effective*** method to obtain two different biased semantic directions in the latent space, that is to set different regularization strength λ.

Appendix C. Additional results on real dataset

In response to the above findings, we show the images edited by different semantic directions, obtained with different regularization strengths λ. The training dataset (sampled from CelebA) is biased where the target attribute $Smiling$ is spuriously correlated with the spurious attributes $Female$ and $Young$. We first use a trained generative model to encode the images into latent codes. Then we train linear classifiers in latent space using logistic regression with different λ. The semantic directions are normal

settings			regularization strength λ				
$n_{maj} : n_{min}$	σ_y	σ_s	1	10	100	1000	10000
2:1	0.1	0.1	0.027	0.032	0.039	0.051	0.072
	0.1	1.0	0.027	0.032	0.040	0.051	0.072
	1.0	0.1	0.026	0.031	0.039	0.051	0.073
	1.0	1.0	0.030	0.033	0.040	0.051	0.073
3:1	0.1	0.1	0.043	0.051	0.063	0.082	0.116
	0.1	1.0	0.043	0.051	0.063	0.082	0.116
	1.0	0.1	0.041	0.050	0.062	0.082	0.117
	1.0	1.0	0.044	0.051	0.063	0.082	0.117
4:1	0.1	0.1	0.054	0.065	0.080	0.104	0.148
	0.1	1.0	0.052	0.063	0.079	0.104	0.148
	1.0	0.1	0.052	0.063	0.079	0.104	0.150
	1.0	1.0	0.055	0.064	0.079	0.104	0.149
5:1	0.1	0.1	0.063	0.076	0.094	0.122	0.175
	0.1	1.0	0.063	0.075	0.093	0.122	0.174
	1.0	0.1	0.061	0.074	0.093	0.122	0.176
	1.0	1.0	0.064	0.075	0.093	0.122	0.175
6:1	0.1	0.1	0.071	0.085	0.105	0.137	0.197
	0.1	1.0	0.070	0.084	0.104	0.137	0.195
	1.0	0.1	0.069	0.083	0.104	0.137	0.199
	1.0	1.0	0.071	0.084	0.104	0.136	0.197
7:1	0.1	0.1	0.077	0.092	0.115	0.150	0.216
	0.1	1.0	0.077	0.092	0.114	0.149	0.214
	1.0	0.1	0.075	0.090	0.113	0.150	0.218
	1.0	1.0	0.077	0.091	0.113	0.149	0.216
8:1	0.1	0.1	0.082	0.099	0.123	0.162	0.233
	0.1	1.0	0.082	0.099	0.122	0.160	0.231
	1.0	0.1	0.080	0.097	0.122	0.161	0.235
	1.0	1.0	0.082	0.097	0.121	0.160	0.233
9:1	0.1	0.1	0.087	0.105	0.131	0.172	0.248
	0.1	1.0	0.087	0.105	0.130	0.171	0.246
	1.0	0.1	0.085	0.103	0.129	0.172	0.250
	1.0	1.0	0.087	0.103	0.129	0.171	0.248
10:1	0.1	0.1	0.092	0.110	0.138	0.181	0.262
	0.1	1.0	0.092	0.110	0.137	0.180	0.260
	1.0	0.1	0.089	0.108	0.136	0.181	0.264
	1.0	1.0	0.092	0.109	0.136	0.180	0.262
11:1	0.1	0.1	0.096	0.115	0.144	0.189	0.275
	0.1	1.0	0.096	0.115	0.143	0.188	0.272
	1.0	0.1	0.093	0.113	0.142	0.189	0.277
	1.0	1.0	0.096	0.114	0.142	0.188	0.275
12:1	0.1	0.1	0.100	0.120	0.150	0.197	0.286
	0.1	1.0	0.099	0.119	0.149	0.196	0.284
	1.0	0.1	0.097	0.118	0.148	0.197	0.289
	1.0	1.0	0.099	0.118	0.148	0.196	0.287
13:1	0.1	0.1	0.103	0.124	0.155	0.205	0.298
	0.1	1.0	0.103	0.124	0.154	0.203	0.294
	1.0	0.1	0.101	0.122	0.154	0.205	0.301
	1.0	1.0	0.103	0.123	0.153	0.203	0.298
14:1	0.1	0.1	0.107	0.128	0.160	0.211	0.308
	0.1	1.0	0.106	0.128	159	0.210	0.306
	1.0	0.1	0.104	0.126	0.159	0.212	0.311
	1.0	1.0	0.106	0.127	0.158	0.210	0.309

Table 6. Results of classifier bias β_{clf} on synthetic dataset. Empirically, in most cases, the classifier bias β_{clf} will be larger, if we increase the regularization strength λ in logistic regression.

vectors of the learned classification boundaries. As shown in Figure 10, a larger λ produces a larger bias in direction, resulting in a more obvious change in spurious attributes.

Appendix D. Implementation Details

For generative modeling, we utilize StyleGAN2 [32] for generator and e4e [67] for encoder. We use HFGI [69] algorithm to train generative models on training dataset with image size of 256 for 30 epochs. The size of features encoded by e4e is (18, 512), and we average over the channels to get latent codes with size of 512. We use regularized logistic regression to obtain directions, and the values of regularization strength λ are 1e+4 and 1e-4 respectively. To get the optimal combination coefficients, we perform grid search and use CLIP [57] as a reference model. More details about combination coefficients are shown in the next subsection. For representation model, we use ResNet-18 [23] for encoder and the representation dimensions are 512. We train the encoder for 135 epochs. We use Adam [36] as optimizer with learning rate 3e-4. We set the editing range [α_l,α_u] as [3,5]. For efficiency, we approximate the sampled degree as an integer. To complete the classification, we fix the encoder and train a linear classifier with Adam until convergence. The learning rate is 1e-2 with 1e-6 weight decay.

Figure 10. Illustration of images edited by different semantic directions, which are trained with different regularization strength λ.