Fair Contrastive Learning for Facial Attribute Classification

Sungho Park[1] Jewook Lee[1] Pilhyeon Lee[1] Sunhee Hwang[2] Dohyung Kim[3] Hyeran Byun[1*]

[1]Yonsei University [2]LG Uplus [3]SK Inc.

✉ qkrtjdgh18@yonsei.ac.kr

Abstract

Learning visual representation of high quality is essential for image classification. Recently, a series of contrastive representation learning methods have achieved preeminent success. Particularly, SupCon [18] outperformed the dominant methods based on cross-entropy loss in representation learning. However, we notice that there could be potential ethical risks in supervised contrastive learning. In this paper, we for the first time analyze unfairness caused by supervised contrastive learning and propose a new Fair Supervised Contrastive Loss (FSCL) for fair visual representation learning. Inheriting the philosophy of supervised contrastive learning, it encourages representation of the same class to be closer to each other than that of different classes, while ensuring fairness by penalizing the inclusion of sensitive attribute information in representation. In addition, we introduce a group-wise normalization to diminish the disparities of intra-group compactness and inter-class separability between demographic groups that arouse unfair classification. Through extensive experiments on CelebA and UTK Face, we validate that the proposed method significantly outperforms SupCon and existing state-of-the-art methods in terms of the trade-off between top-1 accuracy and fairness. Moreover, our method is robust to the intensity of data bias and effectively works in incomplete supervised settings. Our code is available at `https://github.com/sungho-CoolG/FSCL`.

1. Introduction

Learning powerful visual representation is important for reliable performance in image classification. For a long time, most work has relied on cross-entropy loss to learn the representation due to its strong performance [4, 11, 34, 41]. Meanwhile, recent studies based on contrastive learning have been bringing a new paradigm for representation learning [2, 10, 12, 36, 40]. They effectively learn visual representation by drawing positive pairs and pushing away negative ones in the high-dimensional space. Despite being originally introduced for unsupervised learning, the contrastive learning strategy proves to be effective in various vision fields [14, 22, 23]. Particularly, *SupCon* [18] achieves better top-1 accuracy than the state-of-the-art methods based on the cross-entropy loss on ImageNet [32] by simply grafting the contrastive loss to the supervised representation learning.

In this paper, we point out that the contrastive loss may pose potential ethical risks. Despite exhibiting strong performance, it has been underexplored in consideration of fairness which means that the outputs from a model should not be discriminatory in terms of sensitive attributes, such as ethnicity, gender, and age. It is a crucial ethical topic and should be diagnosed in order for the model to be leveraged in the real world [5, 26]. To this end, we analyze the representative contrastive learning model (*SupCon*) from two major perspectives causing unfairness and propose a new contrastive loss to address both of them.

Learning sensitive attribute information is one of the principal causes of unfairness [3, 13, 25, 35]. It incurs unfair classification by inducing a classifier to determine a decision boundary based on undesirable grounds (*i.e.*, sensitive attributes) [19, 28]. From this point of view, we demonstrate that learning sensitive attribute information leads to the decrease of *SupCon* loss on the biased dataset, although the desired behavior is to exclusively learn target class information. Consequently, a model learns both kinds of information to minimize the loss, which eventually aggravates unfairness.

To solve the problem, we propose a Fair Supervised Contrastive Loss (*FSCL*) which prevents encoder networks from learning sensitive attribute information. Basically, it inherits the spirit of supervised contrastive learning that encourages an anchor to be more similar to samples of the same class (*i.e.*, positive samples) than those of other classes (*i.e.*, negative samples). Simultaneously, we limit negative samples to only those having the same sensitive attribute with the anchor among them. In this way, we ensure that learning sensitive attribute information no longer helps the contrastive learning. Rather, it hinders optimizing the loss by increasing the similarity between the anchor and negative samples.

[*]Corresponding author.

On top of that, we analyze *SupCon* in terms of data imbalance between demographic groups, which is another causal factor of unfairness [31]. Concretely, we identify that the imbalanced number of anchors and positive samples between the demographic groups encourages the *SupCon* loss to put more weight on majority groups. As a result, samples from the majority groups generally have higher similarity to the other samples within the same group and lower similarity to samples having different target classes compared to those from the minority groups. We call the former intra-group compactness and the latter inter-class separability. Since their disparities between the groups result in imbalanced classification performances [9, 43, 45], we introduce a group-wise normalization that reduces the gaps by balancing the loss based on the cardinality of anchors and positive samples between the groups. In the experiments, we demonstrate that it further improves fairness with little damage to the classification performance.

To validate the effectiveness of our method, we perform facial attribute classification on CelebA [24] and UTK Face [44] datasets. In various scenarios, the proposed method significantly ameliorates fairness over *SupCon* and outperforms the state-of-the-art methods in terms of the trade-off between classification accuracy and fairness. Besides, our method is robust to the intensity of data bias and effectively improves fairness even in incompletely supervised settings (*e.g.*, without target class labels or with only a few sensitive attribute labels). Furthermore, we show the extensibility of our method to general bias mitigation through experiments on Dogs and Cats dataset [16].

Main contributions. Our main contributions are summarized as follows. 1) We analyze the causes of unfairness in contrastive learning and propose a Fair Supervised Contrastive Loss that improves fairness by penalizing the inclusion of sensitive attribute information in representation. 2) We introduce a group-wise normalization, which mitigates the group-wise disparities of intra-group compactness and inter-class separability that exacerbate unfairness of representation. 3) Through extensive experiments, we validate that our method learns fair representation under various environments. It achieves the best trade-off performances between top-1 accuracy and fairness on CelebA and UTK Face.

2. Related Work

2.1. Fair Representation Learning

Several studies [19, 30, 39, 42] tried to learn fair representation through adversarial training. They adversarially train the encoder network and the classification head for sensitive attributes so that the encoder network is agnostic to sensitive attribute information. [30] learned fair representation by reversing gradients of classification loss for sensitive attributes through gradient reversal layer (GRL). [19] further minimized the mutual information between representation and sensitive attribute labels to eliminate their correlations. [39, 42] designed structures in which the outputs from the classification head for target classes are fed into that for sensitive attributes. The latter head removes bias for sensitive attributes in the intermediate outputs through GRL.

Disentangled representation learning [3, 29, 33] is another mainstream for fair representation learning. [33] enforced two types of representation respectively for target classes and sensitive attributes to be orthogonal to each other by maximizing the entropy of the opposite information in each representation. [3] leveraged the disentanglement loss [20] to separate the representation space into sensitive latents and non-sensitive latents without target class labels. Both methods improved fairness by discarding representation containing sensitive attribute information in downstream classification. Moreover, [29] pointed out the shortcoming of [3] that information related to both target and sensitive attributes is contained in sensitive latents and discarded. They introduced an additional subspace for the intersected information.

Recently, [15] made a fresh attempt to improve fairness without compromising performance through fair knowledge distillation. Based on MMD [7], they encourage the feature distribution of the student model conditioned by sensitive attributes to get close to that of the teacher model averaged over the sensitive attributes. With an oversampling strategy, they ameliorate both classification accuracy and fairness on the balanced test set. Meanwhile, [31] proposed a perturbation method which decorrelates the target and sensitive attributes in the latent space of a pre-trained GAN. Then, they generate a balanced dataset with it and utilized the dataset for a fair training of a classification network.

2.2. Contrastive Representation Learning

Contrastive learning [2, 10, 27, 36, 40] has become a dominant approach to learning visual representation in a self-supervised manner. Without class labels, they learned outstanding representation by pulling samples from the same image together and pushing away those from different images. [27, 36, 40] indicated that the number of negative samples is important for contrastive learning and introduced memory banks to increase it without exploding GPU memory consumption. To solve the inconsistency problem between the updated encoder networks and outdated memory bank, [10] utilized a dynamic memory queue as a memory bank and updated it with a slowly moving momentum encoder. Furthermore, [2] proposed a simple architecture for contrastive learning (*i.e.*, *SimCLR*) that outperforms previous methods without the memory bank and specialized architectures. Based on it, [18] proposed a supervised version of contrastive loss (*i.e.*, *SupCon*). Unlike the previous methods, they set all samples having the same class with an anchor to positive samples and pull them to the anchor.

3. Method

In this section, we first analyze the causes of unfairness in supervised contrastive learning, and then describe the proposed method to solve them. Our method is based on a simple framework for contrastive learning similar to previous works [2, 18]. We note that our key contributions lie in not introducing a specific framework but designing a new general loss for learning fair and informative representation.

3.1. Preliminaries

3.1.1 Overall flow

Assume that we have randomly sampled N data pairs in a batch, $\{x_k, y_k, s_k\}_{k=1...N}$. Here, $x_k \in X$, $y_k \in Y$, and $s_k \in S$ respectively denote an input image, its target class label out of N_y classes, and its sensitive attribute label out of N_s classes. Following the prior works [2, 10], we randomly crop each image x_k to generate two independent patches (*i.e.*, views), $\hat{x}_{2k-1}, \hat{x}_{2k} \in \hat{X}$, resulting in a multi-view batch, $\{\hat{x}_l, \hat{y}_l, \hat{s}_l\}_{l=1...2N}$, where $\hat{y}_{2k-1} = \hat{y}_{2k}$ and $\hat{s}_{2k-1} = \hat{s}_{2k}$ for $k \in [1, N]$. An encoder network $\mathcal{F}(\cdot)$ maps the image patches into representation $H = \{h_l\}_{l=1...2N}$, then a projection network $\mathcal{G}(\cdot)$ in turn maps h_l into another representation $Z = \{z_l\}_{l=1...2N}$ for contrastive learning. The encoding networks (*i.e.*, $\mathcal{F}$ and $\mathcal{G}$) are jointly optimized with contrastive objectives and this process is called representation learning.

After the representation learning process, we freeze the encoder network and throw away the projection network. The frozen encoder network produces representation h_k from the input image x_k instead of the cropped views. Taking the representation as input, a classifier is trained to predict the target class label y_k using the cross-entropy loss.

3.1.2 Self-supervised and supervised contrastive losses

Both self-supervised and supervised contrastive loss enforce an anchor to be more similar to positive samples than negative samples. The major difference is the way positive and negative samples are defined. In the self-supervised version [2], for an anchor $\hat{x}_i$, the other view from the same image is defined as the positive sample. Meanwhile, in the supervised version [18], all patches sharing the same target class labels with the anchor $\hat{x}_i$ are assigned to positive samples, *i.e.*, $\hat{X}_p(i) = \{\hat{x}_p \in \hat{X} | \hat{y}_p = \hat{y}_i, \hat{x}_p \neq \hat{x}_i\}$. In both settings, patches that are neither positive samples nor the anchor are set to negative samples, *i.e.*, $\hat{X}_n(i) = \{\hat{x}_n \in \hat{X} | \hat{x}_n \notin \hat{X}_p(i), \hat{x}_n \neq \hat{x}_i\}$.

In the latent space of z_l, the self-supervised loss maximizes the log-softmax of the similarity between z_i and z_p for the similarity between z_i and representation of all the other samples, $\hat{X}_a(i) = \hat{X}_p(i) \cup \hat{X}_n(i)$. The supervised loss calculates the normalized summation of the multiple log-softmax for all z_p and maximizes it. The self-supervised contrastive loss (L^{SS}) and supervised contrastive loss (L^{Sup}) are formulated as follows.

$$L^{SS} = -\sum_{z_i \in Z} \log \frac{\phi_p}{\sum_{z_a \in Z_a(i)} \phi_a}, \tag{1}$$

$$L^{Sup} = -\sum_{z_i \in Z} \frac{1}{|Z_p(i)|} \sum_{z_p \in Z_p(i)} \log \frac{\phi_p}{\sum_{z_a \in Z_a(i)} \phi_a}, \tag{2}$$

where ϕ_x denotes $exp(z_i \cdot z_x / \tau)$, $x \in \{a, p\}$. τ is a temperature parameter, which is set to lower than 1 for sharper distribution of the softmax scores. $|Z_p(i)|$ is the number of positive samples for an anchor z_i. In L^{Sup}, the cardinality of positive samples varies from anchor to anchor and the factor $\frac{1}{|Z_p(i)|}$ normalizes it. The multiple positive samples and normalization factor ensure that L^{Sup} achieves better classification performances than L^{SS}.

3.2. Unfairness in Supervised Contrastive Loss

3.2.1 Learning of sensitive attribute information

As revealed in the literature [3, 25, 35], learning sensitive attribute information is one of the key factors causing unfair classification. Therefore, to analyze the unfairness of L^{Sup}, we start by exploring whether the loss encourages encoder networks to learn the malignant information.

Specifically, we define learning of sensitive attribute information as increasing $I(Z;S) = \mathbb{E}_{P(z,s)} \log \frac{P(z,s)}{P(z)P(s)}$ [3, 19], which is mutual information between Z and S. Subsequently, we suppose two random points, t_l and t_m, in training time, where $I(Z;S)$ is higher at t_m than at t_l (Assumption 1). In addition, to simplify a wide variety of and complicated data bias, we defined an ideally biased dataset, $\{\tilde{X}, \tilde{Y}, \tilde{S}\}$, where each target attribute is correlated with one different sensitive attribute in equal intensity. We provide further details on it in Appendix. Then we demonstrate that L^{Sup} will lead the encoding networks to learn sensitive attribute information by proving the theorem below.

Theorem 1 Given $\tilde{X}$, $\tilde{Y}$, and $\tilde{S}$, for all $t_l,\ t_m, V^{t_l} > V^{t_m}$.

Here, V^{t_l} and V^{t_m} denote the values of L^{Sup} at t_l and t_m, respectively. Theorem 1 represents that the value of L^{Sup} is always larger at t_l than at t_m. In other words, L^{Sup} is inversely proportional to the $I(Z;S)$. Therefore, it results in the following Corollary. Due to the space limit, we provide the mathematical proof in Appendix.

Corollary 1 Learning sensitive attribute information decreases L_{sup}, given $\tilde{X}$, $\tilde{Y}$, and $\tilde{S}$.

In conclusion, both learning the target attribute and sensitive attribute information reduce L^{Sup}. Since the encoding networks do not have the intrinsic ability to distinguish them, they will learn both kinds of information to optimize L^{Sup}, which eventually aggravates unfairness.

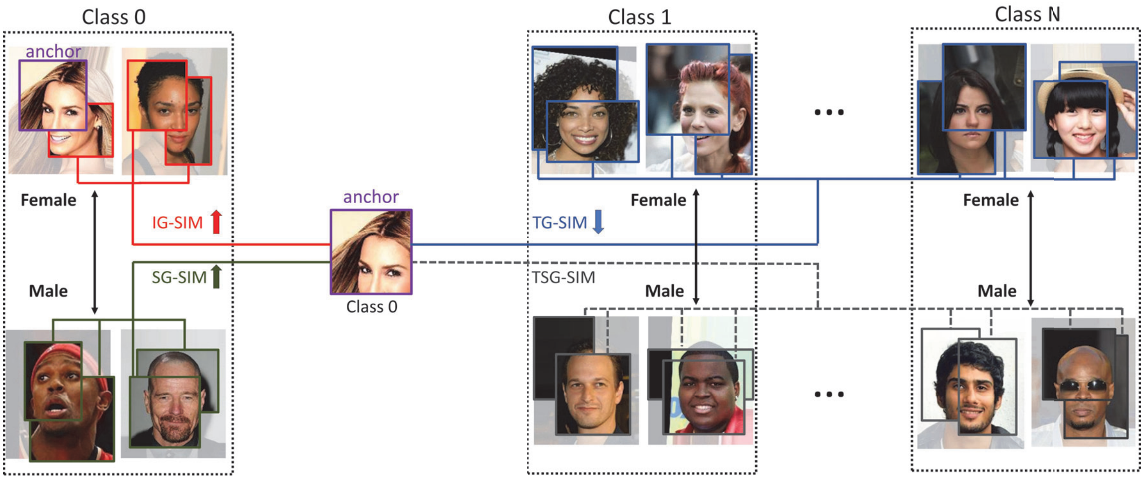

Figure 1. **The concept of fair supervised contrastive loss (*FSCL*).** It increases the similarity between an anchor and samples of the same target class (IG-SIM and SG-SIM). On the contrary, it decreases the similarity between the anchor and samples having different target classes and the same sensitive attribute, *e.g.,* gender, (TG-SIM). Here TSG-SIM is not directly affected by the loss.

3.2.2 Data imbalance between demographic groups

In addition to discovering the malignant information learning, we explore another cause of unfairness with respect to the imbalanced number of data between data groups. Concretely, we define a data group as a bundle of data having common target classes and sensitive attributes. Based on the group definition, we reformulate L^{Sup} as follows.

$$L^{Sup} = -\sum_{\forall j,k} \sum_{z_i \in Z^{j,k}} \frac{1}{|Z_p(i)|} \sum_{\forall k} \sum_{z_p \in Z_p^k(i)} \log \frac{\phi_p}{\sum_{z_a \in Z_a(i)} \phi_a}, \tag{3}$$

where $Z^{j,k} = \{z_i \in Z | \hat{y}_i = j, \hat{s}_i = k\}$ and $Z_p^k(i) = \{z_p \in Z_p(i) | \hat{s}_p = k\}$ for $j \in [1, N_y]$, and $k \in [1, N_s]$. We note that Eq. 2 and 3 are mathematically identical, but the latter reveals that the imbalanced number of anchors between $Z^{j,k}$ and that of positive samples between $Z_p^k(i)$ are not normalized by the existing factor $1/|Z_p(i)|$. It results in the loss putting more weight on the majority groups, and thus the loss encourages the majority groups to have better intra-group compactness and inter-class separability compared to the minority groups. Consequently, as indicated in [9, 43, 45], their disparities between the groups incur unfair classification performances.

3.3. Fair Supervised Contrastive Loss

To resolve the problem of learning sensitive attribute information (Sec. 3.2.1), we propose a Fair Supervised Contrastive Loss (*FSCL*) which explicitly penalizes that the encoding networks learn the unwanted information. For brief and clear explanations, we define the following similarities in consideration of the relationship between an anchor and the other samples.

- **IG-SIM** (Intra-Group Similarity) is the similarity between an anchor and samples within the same group, which have the *same* target class and the *same* sensitive attribute with the anchor. The sample set is defined as $Z_{ig}(i) = \{z_{ig} \in Z_p(i) | \hat{s}_{ig} = \hat{s}_i, \hat{y}_{ig} = \hat{y}_i\}$.
- **SG-SIM** (Sensitive Inter-Group Similarity) is the similarity between an anchor and samples that belong to groups with the *same* target class and *different* sensitive attributes with the anchor. The sample set is defined as $Z_{sg}(i) = \{z_{sg} \in Z_p(i) | \hat{s}_{sg} \neq \hat{s}_i, \hat{y}_{sg} = \hat{y}_i\}$.
- **TG-SIM** (Target Inter-Group Similarity) is the similarity between an anchor and samples that belong to groups with *different* target classes and the *same* sensitive attribute with the anchor. The sample set is defined as $Z_{tg}(i) = \{z_{tg} \in Z_n(i) | \hat{s}_{tg} = \hat{s}_i, \hat{y}_{tg} \neq \hat{y}_i\})$
- **TSG-SIM** (Target & Sensitive Inter-Group Similarity) is the similarity between an anchor and samples that belong to groups with *different* target classes and *different* sensitive attributes with the anchor. The sample set is defined as *i.e.*, $Z_{tsg}(i) = \{z_{tsg} \in Z_n(i) | \hat{s}_{tsg} \neq \hat{s}_i, \hat{y}_{tsg} \neq \hat{y}_i\}$.

Our key idea is to define the negative sample set as the samples with the same sensitive attributes and different target classes with the anchor (*i.e.*, Z_{tg}). Based on this, we design *FSCL* that encourages IG-SIM and SG-SIM to be higher than TG-SIM, as illustrated in Figure 1. It is formulated as follows.

$$FSCL = -\sum_{z_i \in Z} \frac{1}{|Z_p(i)|} \sum_{z_p \in Z_p(i)} \log \frac{\phi_p}{\sum_{z_{tg} \in Z_{tg}(i)} \phi_{tg}}, \quad (4)$$

where $Z_p(i) = Z_{ig}(i) \cup Z_{sg}(i)$ and $\phi_{tg} = exp(z_i \cdot z_{tg}/\tau)$.

On a case-by-case basis, we explain how our *FSCL* addresses the problem of learning unwanted information. In a case of $z_p \in Z_{ig}(i)$, the positive samples and negative samples (*i.e.*, $Z_{tg}(i)$) all have the same sensitive attributes with the anchor. Therefore, the encoding networks no longer consider the sensitive attribute information to be a valuable feature for contrasting an anchor with the negative samples more than with the positive samples.

In the other case of $z_p \in Z_{sg}(i)$, the positive samples have different sensitive attributes from the anchor and negative samples (*i.e.*, $Z_{tg}(i)$). If the encoding networks learn the sensitive attribute information, the similarity between the positive samples and the anchor (*i.e.*, ϕ_p) will decrease and the similarity between the negative samples and it (*i.e.*, ϕ_{tg}) will increase, which is contrary to the objective of the loss. As a result, minimizing the loss inhibits learning the unwanted information in this case.

3.4. Group-wise Normalization

As aforementioned in Sec. 3.2.2, the imbalanced number of anchors and positive samples between data groups causes the group-wise disparities in terms of intra-group compactness and inter-class separability. To alleviate the unfairness brought by the disparities, we introduce group-wise normalization as follows.

$$FSCL\text{+} = -\sum_{\forall j,k} \frac{1}{|Z^{j,k}|} \sum_{z_i \in Z^{j,k}} \sum_{\forall k} \frac{1}{|Z_p^k(i)|} \sum_{z_p \in Z_p^k(i)} \log \frac{\phi_p}{\sum_{z_{tg} \in Z_{tg}(i)} \phi_{tg}}, \quad (5)$$

where $1/|Z^{j,k}|$ and $1/|Z_p^k(i)|$ are the group-wise normalization factors. Different from the existing factor in L^{Sup} (*i.e.,* $1/|Z_p(i)|$), they normalize the cardinality of anchors and positive samples within each group. On an experimental basis, we demonstrate that the proposed normalization mitigates the group-wise imbalances in terms of intra-group compactness and inter-class separability.

4. Experiment

4.1. Datasets

CelebA [24] contains about 200k facial images with 40 binary attribute annotations. We set *male* (*m*) and *young* (*y*) to sensitive attributes and select target attributes having the highest Pearson correlation with the sensitive attributes [1, 37]. Amongst, we manually excluded the extremely correlated attributes for reliable evaluation. For heavy-makeup as example, there are only 22 males with heavy-makeup in test set. As a result, we exploit three single target attributes: *attractiveness* (*a*), *bignose* (*b*), and *bags-under-eyes* (*e*) as well as two pairs of target attributes: {*bignose*, *bags-under-eyes*} and {*attractiveness*, *mouth-slightly-open* (*o*)}.

UTK Face [44] consists of about 20k facial images with three kinds of annotations: *gender*, *age*, and *ethnicity*. To evaluate fairness in varied levels of data imbalance, we design several imbalanced versions for the training set. Note that the standard protocol on data splits is not provided in this dataset. Concretely, we set *age* and *ethnicity* to the sensitive attributes and *gender* to the target attribute. *Age* and *ethnicity* are reformed to binary attributes based on whether *age* is under 35 or not and *ethnicity* is Caucasian or not, respectively. A sensitive group (*e.g.*, Caucasian) has male data α times as much as female data and the other sensitive group has the opposite gender ratio. α is set to 2, 3, and 4 to simulate varying bias levels. Unlike the training set, we organize completely balanced validation and test sets for a fair evaluation.

Dogs and Cats [16] has 38,500 dog or cat images. In addition to the original species labels (dog or cat), LNL [19] further annotated color labels (bright or dark). We set *color* to the sensitive attribute and *species* to the target attribute. We compose a *color* biased training set that contains 5 times more black cats than white cats, while 5 times more white dogs than black dogs. For a fair evaluation, we compose the test set to be completely balanced. Note that we utilize this dataset to examine the extensibility of the proposed method to general bias mitigation (*i.e.*, color) beyond its fairness.

We provide more details on the datasets in Appendix.

4.2. Fairness Metrics

A variety of fairness notions are exploited to measure fairness in classification tasks (*e.g.*, demographic parity [21], equal opportunity, and equalized odds [8]). Demographic parity means that the proportion of positive outcomes in each sensitive group should be equal. Although it may be used as reliable metrics in situations where equality of outcome has to be guaranteed, there is a drawback in that a classifier should deliberately misclassify some labels to satisfy it if the proportion of positive outcomes is not equal in the ground truth (GT) [6, 8]. Equal opportunity solves this issue by pursuing the equal true positive rate (TPR) between sensitive groups. However, it does not address unfairness in negative outcomes. In many real-world applications such as facial attribute classification, fairness of positive and negative outcomes is equivalently important. Therefore, equalized odds, which demands both the equal TPR and false positive rate (FPR), are the most suitable to measure fairness in our experiments. Following the definition in [8], we measure the degree of equalized odds (EO) in various settings (*e.g.*, mul-

Method	T=*a* / S=*m*		T=*a* / S=*y*		T=*b* / S=*m*		T=*b* / S=*y*		T=*e* / S=*m*		T=*e* / S=*y*		T=*a* & *o* / S=*m*		T=*e* & *b* / S=*m*		T=*a* / S=*m* & *y*	
	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.	EO	Acc.
CE [11]	27.8	79.6	16.8	79.8	17.6	84.0	14.7	84.5	15.0	83.9	12.7	83.8	30.0	73.9	12.9	72.6	31.3	79.5
GRL [30]	24.9	77.2	14.7	74.6	14.0	82.5	10.0	83.3	6.7	81.9	5.9	82.3	17.8	73.1	9.4	71.4	22.9	78.6
LNL [19]	21.8	79.9	13.7	74.3	10.7	82.3	6.8	82.3	5.0	81.6	3.3	80.3	16.7	72.9	7.4	70.8	20.7	77.7
FD-VAE [29]	15.1	76.9	14.8	77.5	11.2	81.6	6.7	81.7	5.7	82.6	6.2	84.0	18.2	73.4	8.2	70.2	19.9	78.0
MFD [15]	7.4	78.0	14.9	80.0	7.3	78.0	5.4	78.0	8.7	79.0	5.2	78.0	8.7	74.0	9.0	70.0	19.4	76.1
SupCon [18]	30.5	80.5	21.7	80.1	20.7	84.6	16.9	84.4	20.8	84.3	10.8	84.0	22.8	74.0	12.5	72.7	24.4	81.7
FSCL	11.5	79.1	13.0	79.1	7.0	82.1	6.4	83.8	3.8	82.7	1.8	82.0	8.1	74.1	6.8	71.1	19.9	79.4
FSCL+	**6.5**	79.1	**12.4**	79.1	**4.7**	82.9	**4.8**	84.1	**3.0**	83.4	**1.6**	83.5	**3.6**	74.8	**2.5**	70.8	**17.0**	77.2

Table 1. **Classification results on CelebA.** We measure classification accuracy (ACC.) and equalized odds (EO) in various scenarios. Here *a*, *b*, *e*, *o*, *m*, and *y* respectively denote *attractiveness*, *bignose*, *bags-under-eyes*, *mouth-slightly-open*, *male*, and *young*. On the other hand, T and S represent target and sensitive attributes, respectively. All the results are the averaged scores over three independent runs. The standard deviations are provided in Appendix.

tiple classes or sensitive attributes) as follows.

$$\overline{\sum_{\forall y, c, \{s^0, s^1\} \subset S}} \left| P_{s^0}(C = c \,|Y = y) - P_{s^1}(C = c \,|Y = y) \right|, \tag{6}$$

where $\overline{\sum}$ is the averaged sum. $y \in Y$ and $c \in C$ are target labels and outputs from a classifier, respectively, and $\{s^0, s^1\}$ is a two-element subset of sensitive attribute groups S.

4.3. Implementation Details

For contrastive learning, we utilize ResNet-18 [11] for the encoder network $\mathcal{F}$ and a MLP with two hidden layers for the projection network $\mathcal{G}$. The dimensions of latent spaces are set to 256 and 128, respectively. We augment two cropped patches per image following the augmentation strategy in [2] and resize them to 128×128. We set the temperature parameter τ to 0.1 based on the analysis in [18]. We train the encoding networks for 100 epochs in the representation learning stage, and subsequently train the classifier, which is a MLP with one hidden layer, for 10 epochs using the cross-entropy loss. For the experiments with multiple target or sensitive attributes, we combine multiple binary attribute labels into a multi-class label. All comparative models share the same structures of the encoder network and classifier as ours for a fair comparison. The results reported in this paper are averaged over three independent runs. More details for the augmentation strategy, structure of networks, and experiment settings are provided in Appendix.

4.4. Classification Results on CelebA

Table 1 shows the classification results on CelebA. For diverse combinations of target and sensitive attributes, we measure classification performances and fairness with top-1 accuracy and equalized odds (EO), respectively. In all the experiments, *Cross-Entropy* (CE) [11] and *SupCon* [18] record excellent top-1 accuracy but suffer from severe unfairness. Notably, the proposed methods (*FSCL* and *FSCL+*) significantly improve EO over them while preserving the competitive performances. Particularly, the comparison between *FSCL* (blue) and *FSCL+* (red) shows that the group-wise normalization brings about better fairness while well preserving the performance or even improving it. Furthermore, we compare ours with various state-of-the-art approaches for fairness such as adversarial training (*GRL* [30] and *LNL* [19]), disentangled representation learning (*FD-VAE* [29]), and fair distillation (*MFD* [15]). *FSCL+* substantially outperforms all the state-of-the-art methods in terms of the trade-off between top-1 accuracy and EO in all the settings. For a clearer comparison of the trade-off performances, we also provide the experimental results in figure form in Appendix.

Method	Adversarial Training [30]	EO (↓)	Acc. (↑)
SupCon [18]	✗	$30.5_{\pm 1.3}$	$80.5_{\pm 0.7}$
	✓	$21.0_{\pm 0.5}$	$76.6_{\pm 0.3}$
FSCL+	✗	$\mathbf{6.5}_{\pm 0.4}$	$79.1_{\pm 0.1}$
	✓	$9.0_{\pm 0.5}$	$79.2_{\pm 0.1}$

Table 2. **Effect of adversarial training in contrastive learning on CelebA dataset.** We utilize GRL [30] for the adversarial training. Here $attractiveness$ and $male$ are set to the target class and sensitive attribute, respectively.

4.5. Adversarial Training in Contrastive Learning

Intuitively, to mitigate the unfairness of *SupCon*, one may imagine simply combining adversarial training with it. In Table 2, we demonstrate the effect of adversarial training by applying *GRL* [30] to *SupCon* and *FSCL+* in the representation learning. For *SupCon*, while improving fairness to an extent, it largely damages the classification performance. In addition, *FSCL+* achieves much better EO and top-1 accuracy than *SupCon* combined with *GRL*. This indicates that the simple graft of adversarial training to contrastive learning does not sufficiently improve fairness and designing a new method seamlessly integrated into contrastive learning is more effective. We do not see further improvements in EO when applying *GRL* to FSCL+.

Method	Ramaswamy *et al.* [31]	EO (↓)	Acc. (↑)
Cross-Entropy [11]	✗	$27.8_{\pm 0.2}$	$79.6_{\pm 0.5}$
	✓	$24.1_{\pm 0.5}$	$79.6_{\pm 0.2}$
FSCL+	✗	$6.5_{\pm 0.4}$	$79.1_{\pm 0.1}$
	✓	$\mathbf{4.2}_{\pm 0.4}$	$79.6_{\pm 0.1}$

Table 3. **Compatibility with fair data augmentation [31] on CelebA dataset**. We set *attractiveness* and *male* to the target class and sensitive attribute, respectively.

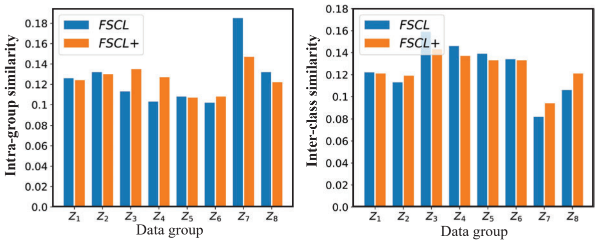

Figure 2. **Effectiveness of group-wise normalization.** The group-wise normalization (*i.e.*, *FSCL+*) significantly mitigates the group-wise disparities in terms of intra-group and inter-class similarities.

4.6. Compatibility with Fair Data Augmentation

We incorporate our method with Ramaswamy *et al.* [31], one of the state-of-the-art pre-processing methods for fair classification. It generates a de-biased dataset through a Progressive GAN [17] and augments the original dataset with the generated one. In Table 3, we report the performances of the baseline (*i.e., Cross-Entropy*) and *FSCL+* trained on the original/augmented dataset. The results show that *FSCL+* outperforms Ramaswamy *et al.* (2nd row) in terms of both EO and top-1 accuracy. Besides, the fairness of ours is further enhanced when adopting the fair data augmentation, which indicates its compatibility.

4.7. Effectiveness of Group-wise Normalization

To analyze the effectiveness of the group-wise normalization, we compare intra-group compactness and inter-class separability between *FSCL* and *FSCL+*. To this end, we first divide the test set into 8 groups with respect to one target class, *attractiveness*, and two sensitive attributes, *male* and *young*, and then calculate them as follows. The former is measured by averaging the similarities between representation within a group (*i.e.*, intra-group similarity) and the latter is measured by averaging the similarity between representation in a group and representation having different class labels with it (*i.e.*, inter-class similarity). For easier comparison, the values are normalized to sum to unity, as shown in Figure 2. The plots demonstrate that the group-wise normalization significantly diminishes the group-wise disparities. In specific, *FSCL* has the standard deviations of 0.084 and 0.031 in intra-group and inter-class similarities, respectively, while *FSCL+* has lower standard deviations of 0.038 and 0.024.

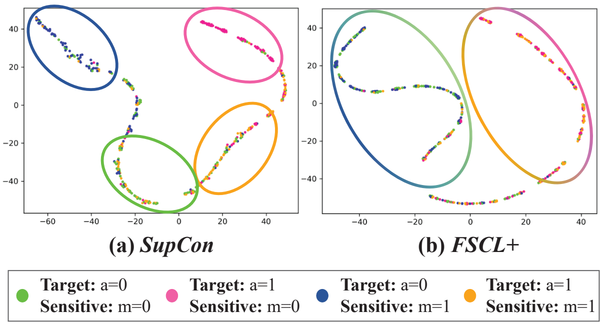

Figure 3. **Qualitative comparison using t-SNE visualizations.** It is clearly shown that *FSCL+* (b) learns representation more independent of the sensitive attribute than *SupCon* (a).

4.8. Qualitative Analysis with t-SNE Visualization

In Figure 3, we provide t-SNE plots [38] of representation from *SupCon* and *FSCL+* on CelebA dataset. The representation is divided into 4 groups in terms of the target class (*i.e.*, *attractiveness*) and sensitive attribute (*i.e.*, *male*), which are visualized in different colors. In *SupCon*, the representation is divided by both the target class and sensitive attribute, suggesting that the encoding networks learn information for the sensitive attribute as well as the target class. Consequently, the representation of minority groups (*i.e.*, green and orange colors) is more similar to the representation of the counterpart class than that of majority groups (*i.e.*, blue and pink colors). In contrast, in *FSCL+*, the representation is divided by only the target class, that is, it is more agnostic to the sensitive attribute. Accordingly, majority groups can no longer have more privileges than minority groups, which explains why our loss can achieve fairer performance than *SupCon* in image classification. Details of experimental settings are provided in Appendix.

4.9. Robustness to Severity of Data Bias

In Figure 4, we present the trends of EO and top-1 accuracy according to the intensity shift of data imbalance (α) on UTK Face dataset. It can be clearly noticed that our loss best prevents the degradation of fairness caused by an increase in α, achieving the fairest performance at all the intensities. In the figure, as α increases, the EO gaps between ours and the others become larger, which manifests the robustness of the proposed methods against the severity of data bias. Moreover, at all the intensities, our loss successfully maintains the top-1 accuracy, which is close to *SupCon*. Experimental results on another sensitive attribute (*i.e.*, *age*) draw similar conclusions and are provided in Appendix.

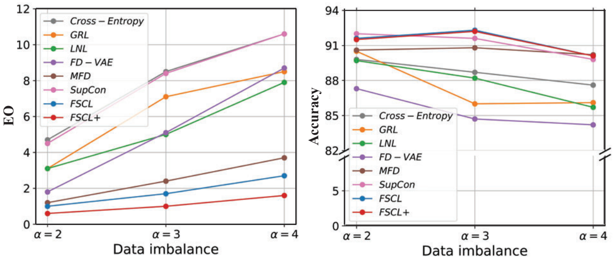

Figure 4. **Classification results on UTK Face.** We measure classification accuracy and equalized odds (EO) under different severity levels of imbalance (α). We set *gender* and *ethnicity* to the target class and sensitive attribute, respectively. Larger α indicates that the training set is more imbalanced.

Method	Target:a / Sensitive:m		Target:b / Sensitive:m	
	EO (↓)	Acc. (↑)	EO (↓)	Acc. (↑)
SimCLR [2]	$29.4_{\pm 2.5}$	$75.7_{\pm 0.2}$	$16.4_{\pm 0.4}$	$82.0_{\pm 0.1}$
SimCLR [2] + *GRL* [30]	$21.9_{\pm 0.9}$	$72.3_{\pm 0.4}$	$13.7_{\pm 0.3}$	$82.3_{\pm 0.0}$
FSCL†	$\mathbf{14.8}_{\pm 0.9}$	$74.6_{\pm 0.4}$	$\mathbf{6.1}_{\pm 0.6}$	$80.8_{\pm 0.2}$

Table 4. **Classification results on CelebA in the absence of target class labels during representation learning.** *FSCL*† is a modified version of *FSCL* that does not use the target class labels.

4.10. Results in Incomplete Supervised Settings

We explore a more challenging problem setting, where target class labels are unavailable during the representation learning process. To this end, we introduce a modified version of *FSCL* that does not exploit target class labels, which is denoted by *FSCL*†. Similar to *SimCLR* [2], it uses only a single positive sample that comes from the same image with an anchor. As shown in Table 4, ours significantly improves fairness at the acceptable cost of top-1 accuracy, compared to *SimCLR* and *SimCLR+GRL*. Details of the modification are provided in Appendix.

Moreover, we conduct experiments under another challenging environment where only a small portion of data have sensitive attribute labels. One of our simple strategies to handle this task is to generate pseudo-labels for applying *FSCL+* loss. Specifically, we train a classifier to predict sensitive attribute labels only with the samples having sensitive attribute labels, and then generate the pseudo labels of sensitive attributes for the other samples. Another strategy is to apply *FSCL+* loss only to data with sensitive labels and *SupCon* to the other data. Table 5 shows that *FSCL+* effectively ameliorates EO over *SupCon* even under the incomplete supervision of sensitive attributes. Surprisingly, *FSCL+* with only 5% of labels is able to outperform *SupCon+GRL* using all the labels.

4.11. Extensibility to General Bias Mitigation

To verify the efficacy of the proposed methods in a general bias type, we conduct experiments on Dogs and Cats [16] with *color* bias. The results are shown in Figure 5, where our models (*FSCL* and *FSCL+*) best eliminate the color bias, which implies that they are generalizable to various types of bias. Note that *FSCL*, *FSCL+*, and *MFD* show higher top-1 accuracy than their baselines since fairness improves the performance on the completely balanced test set.

Method	# of Sensitive	Pseudo-labeling	EO (↓)	Acc. (↑)
SupCon [18]	0	-	$30.5_{\pm 1.3}$	$80.5_{\pm 0.7}$
SupCon [18] + *GRL* [30]	1	-	$21.0_{\pm 0.5}$	$76.6_{\pm 0.3}$
FSCL+	1	-	$\mathbf{6.5}_{\pm 0.4}$	$79.1_{\pm 0.1}$
	1/2	✗	$13.4_{\pm 0.1}$	$79.3_{\pm 0.3}$
		✓	$12.8_{\pm 1.2}$	$79.4_{\pm 0.3}$
	1/4	✗	$18.7_{\pm 0.3}$	$80.0_{\pm 0.3}$
		✓	$13.4_{\pm 0.1}$	$79.5_{\pm 0.5}$
	1/10	✗	$20.7_{\pm 0.5}$	$80.2_{\pm 0.1}$
		✓	$16.5_{\pm 0.5}$	$79.6_{\pm 0.4}$
	1/20	✗	$23.4_{\pm 0.0}$	$80.6_{\pm 0.1}$
		✓	$18.8_{\pm 1.1}$	$78.5_{\pm 0.2}$

Table 5. **Classification results on CelebA under incomplete supervision of sensitive attribute labels.** "# of Sensitive" denotes the ratio of data having sensitive attribute labels. We set *attractiveness* and *male* to the target class and sensitive attribute, respectively.

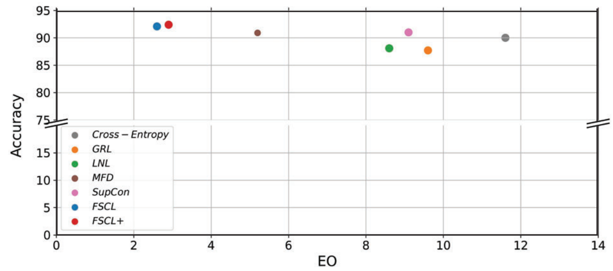

Figure 5. **Classification results on Dogs and Cats.** We set *species* and *color* to the target and sensitive attributes, respectively.

5. Conclusion

In this paper, we addressed fairness in contrastive learning. We first analyzed the causative factors of unfairness in the supervised contrastive loss. Then we proposed the fair supervised contrastive loss and introduced the group-wise normalization into the loss. Through extensive experiments, we validated that our loss effectively improves fairness with little degradation of the classification performance.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2B5B02001467) and Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01396: Development of Framework for Analyzing, Detecting, Mitigating of Bias in AI model and Training Data, No. 2020-0-01361: Artificial Intelligence Graduate School Program (Yonsei University)).

References

[1] Luigi Celona, S. Bianco, and R. Schettini. Fine-grained face annotation using deep multi-task cnn. *Sensors (Basel, Switzerland)*, 18, 2018. 5

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 1, 2, 3, 6, 8

[3] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1, 2, 3

[4] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. 1

[5] Conor Dougherty. Google photos mistakenly labels black people gorillas. Twitter, 2015. 1

[6] 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*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. 5

[7] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773, mar 2012. 2

[8] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. 5

[9] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6468–6478, 2019. 2, 4

[10] 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 (CVPR)*, June 2020. 1, 2, 3

[11] 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. 1, 6, 7

[12] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. 1

[13] Sunhee Hwang, Sungho Park, Pilhyeon Lee, Seogkyu Jeon, Dohyung Kim, and Hyeran Byun. Exploiting transferable knowledge for fairness-aware image classification. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 1

[14] Seogkyu Jeon, Kibeom Hong, Pilhyeon Lee, Jewook Lee, and Hyeran Byun. Feature stylization and domain-aware contrastive learning for domain generalization. In *The 29th ACM International Conference on Multimedia*, pages 22–31, 2021. 1

[15] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12115–12124, June 2021. 2, 6

[16] Kaggle. Dogs vs. cats. 2013. 2, 5, 8

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 7

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 1, 2, 3, 6, 8

[19] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 3, 5, 6

[20] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 2

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017. 5

[22] Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13648–13657, 2021. 1

[23] Pilhyeon Lee, Sunhee Hwang, Jewook Lee, Minjung Shin, Seogkyu Jeon, and Hyeran Byun. Inter-subject contrastive learning for subject adaptive eeg-based visual recognition. In *The 10th International Winter Conference on Brain-Computer Interface (BCI)*, 2022. 1

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of Inter-*

national Conference on Computer Vision (ICCV)*, December 2015. 2, 5

[25] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations, 2019. 1, 3

[26] Natasha Lomas. Faceapp apologizes for building a racist ai. TechCrunch, 2018. 1

[27] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[28] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20673–20684. Curran Associates, Inc., 2020. 1

[29] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. *Proceedings of AAAI-2021*, 2021. 2, 6

[30] Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198, 2018. 2, 6, 8

[31] Vikram V. Ramaswamy, Sunnie S. Y. 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 (CVPR)*, pages 9301–9310, June 2021. 2, 7

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[33] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 746–761. Springer, 2020. 2

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[35] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173. PMLR, 16–18 Apr 2019. 1, 3

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 776–794, Cham, 2020. Springer International Publishing. 1, 2

[37] Robert Torfason, Eirikur Agustsson, Rasmus Rothe, and Radu Timofte. From face images and attributes to attributes. In *Asian Conference on Computer Vision*, pages 313–329. Springer, 2016. 5

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, June 2015. 7

[39] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018. 2

[40] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[41] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[42] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery. 2

[43] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5419–5428, 2017. 2, 4

[44] 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 (CVPR)*, July 2017. 2, 5

[45] He Zhu and Shan Yu. Intra-class uncertainty loss function for classification. *CoRR*, abs/2104.05298, 2021. 2, 4