PRECAUTIONARY UNFAIRNESS IN SELF-SUPERVISED CONTRASTIVE PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, self-supervised contrastive pre-training has become the *de facto* regime, that allows for efficient downstream fine-tuning. Meanwhile, its fairness issues are barely studied, though they have drawn great attention from the machine learning community, where structured biases in data can lead to biased predictions against under-presented groups. Most existing fairness metrics and algorithms focus on supervised settings, e.g., based on disparities in prediction performance, and they become inapplicable in the absence of supervision. We are thus interested in the challenging question: *how does the pre-training representation (un)fairness transfer to the downstream task (un)fairness, and can we define and pursue fairness in unsupervised pre-training?* Firstly, we empirically show that imbalanced groups in the pre-training data indeed lead to unfairness in the pre-trained representations, and that cannot be easily fixed by fairness-aware fine-tuning without sacrificing efficiency. Secondly, motivated by the observation that the majority group of the pre-training data dominates the learned representations, we design the first unfairness metric that can be applicable to self-supervised learning, and leverage that to guide the contrastive pre-training for fairness-aware representations. Our experiments demonstrate that the underestimated representation disparities strike over 10% surges on the proposed metric and our algorithm improves 10 out of 13 tasks on the 1%-labeled CelebA dataset. Codes will be released upon acceptance.

1 INTRODUCTION

Supervised learning has achieved remarkable success in a variety of fields thanks to the availability of massive amount of data and the flexibility of deep models. On the other hand, in many applications the data annotation process remains expensive and therefore we cannot directly apply supervised learning. To mitigate this issue, recently many studies investigated self-supervised approaches to leverage label-free data for pre-training, and learn generalizable representations for downstream learning tasks. A prominent example is contrastive learning (CL) (Goyal et al., 2019), which trains between-sample discriminative and transformation invariant representations by self-created supervisions. Outstandingly, when combined with an efficient linearly fine-tuning (LFT) over downstream tasks, the overall predictive performance surpasses state-of-the-art supervised learning in various applications (Chen et al., 2020a; Grill et al., 2020; Tian et al., 2020; Wang & Isola, 2020; Chen et al., 2020c).

Despite the impressive downstream predictive performance, the potential representation bias introduced by CL datasets is barely studied. Especially when different social groups are involved, structural biases in data or representations can induce biases in the learned models that are unfair to under-represented groups. One such example is that people may be predicted as criminals by models due to their skin colors, simply because of statistical racial bias on skin colors in data (Redmond & Baveja, 2002). A recent theoretical result in (Dutta et al., 2020) established the connection between the concerned unfairness and a bad feature space: the group-disparate class separability (the optimal discrimination in the space) results in an inevitable sacrifice of fairness for accuracy, no matter what kind of fairness mitigation is applied. The connection suggests that once unfairness is introduced in the pre-training stage, the cost to mitigate it during the fine-tuning stage can be prohibitive. We therefore ask the following questions:

(1) Will the pre-training representation unfairness result in downstream unfairness? And if so, (2) how can we measure and mitigate it early in pre-training?

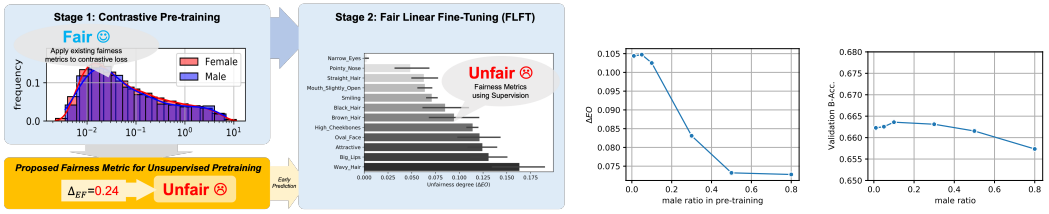

Figure 1: Illustration of potential downstream unfairness induced by biased data in pre-training, and the relationship between unfairness/balanced-accuracy (B-Acc) and pre-training group imbalance. The unfairness retains even after fair fine-tuning. Unlike the downstream supervised learning, pre-training does not exhibit a group-biased loss distribution, which introduces challenges for loss-based unfairness detection and results in a false sense of fairness. In contrast, our proposed metric Δ_{FD} is feature-based and can successfully predict downstream unfairness (Δ_{EO}) as early as in the pre-training stage.

Our work is the first attempt to address the above questions by a rigorous and comprehensive study. For the first question, we look for the biases in CL representations that will impact the downstream fairness even under fairness protected LFT. We show that the biased data in the pre-training stage induce unfair representations, which carry the unfairness to greatly impact the downstream task. The fairness issues induced by unfair representations cannot be fixed in downstream tasks, even by fairness mitigation.

With this positive answer to Q1, it is still challenging to answer the second question, because of the inconsistent learning goals of CL and supervised LFT. Most existing fairness metrics used disparity of task losses (or similar accuracy variants) between groups, and therefore cannot be applied in the pre-training phase, which is task independent. To see such discrepancy, a "hard" sample characterized by a large loss by the CL stage can have a small classification loss in the LFT stage, and thus disparity as shown in CL losses may NOT lead to unfairness as measured by LFT losses. For example, given a cluster of similar samples, they could be of small LFT losses because they likely belong to the identical downstream class, while of large CL losses because the dissimilarity among them is large. As such, directly applying existing fairness metrics and algorithms on the CL phase to calibrate its loss distributions may lead to a false sense of fairness, as illustrated in Fig. 1.

To tackle this problem, we identify and study an important source of downstream unfairness: *the imbalance of learned features* during pre-training, i.e., the majority group activated more discriminative features than the minority group. As an example, because that people with a darker skin color (minority) can be easily differentiated from those of lighter skin color (majority) by their skin colors, the representation learning via CL will more focus on such group-discriminative features, and other features are potentially task-discriminative will be ignored. Such learned representations will lead to worse downstream performance for minority. As shown in Fig. 1, our fairness metric Δ_{FD} can recognize the downstream unfairness (measured by a widely adopted metric Δ_{EO}) early in pre-training.

To conclude, our main contributions are as follows. **(1)** For the first time, we conducted a comprehensive investigation on the unfairness transferred from contrastive pre-training to fine-tuning, and showed that such unfairness cannot be fixed by linear fair fine-tuning. **(2)** We defined a novel feature-based fairness metric for task-agnostic CL, *Equalized Feature (EF)*, which is shown to characterize fairness in downstream tasks (e.g., correlate with equalized odds). **(3)** Following the principle of EF, we proposed a novel fair representation learning for CL by augmenting the features of the under-presented group by a simple yet efficient feature masking strategy. **(4)** Through extensive experiments, we showed that our method effectively improves fairness in multiple downstream tasks.

2 Related Work

Supervised unfairness mitigation aims to reduce the prediction disparity between two protected social groups (Luong et al., 2011; Kamishima et al., 2011), for example, higher predicted crime probability for one race than other races. It can be categorized into three classes according to the execution periods. **(1) Pre-processing** methods reduce the group discrimination in datasets by sampling or re-weighting (Calders et al., 2009; Kamiran & Calders, 2012; Calmon et al., 2017), or

learning disentangled representations (Zemel et al., 2013). **(2) In-processing** methods secure fairness by introducing constraints that limit correlation between sensitive attributes and labels (Zafar et al., 2017a;b; Donini et al., 2020), or fairness inequality by a two-player game (Cotter et al., 2019; Donini et al., 2020; Komiyama et al., 2018). **(3) Post-processing** methods adjust trained model to ensure fairness (Hardt et al., 2016; Kim et al., 2019), but are shown to be suboptimal w.r.t. the notion of equalized odds (Woodworth et al., 2017). Most of these approaches rely on the existence of prediction labels to measure and ensure fairness. This paper considers a novel problem setting that addresses unfairness without labels.

Self-supervised contrastive learning. Recently, contrastive learning (CL) (Chen et al., 2020a;c; van den Oord et al., 2019; He et al., 2020; Chen et al., 2020b) and its variants (Grill et al., 2020; Tian et al., 2020; Chen & He, 2021) are powerful self-supervised representation learning approaches without requiring labels. Given a unlabeled sample, CL and its variants train a feature encoder by maximizing similarity between positive samples (transformations of the same sample) and dissimilarity between negative ones (transformation of different samples). CL uniformly improves local discrimination among all samples (Wang & Isola, 2020; Wang & Liu, 2021), and is not biased toward group-wise or class-wise discrimination. Perhaps due to the property, the learned feature representations improve the downstream fairness (Ramapuram et al., 2021; Pruksachatkun et al., 2021). However, it remains unclear when CL learns group-fair representations. Our work shows that CL can suffer from imbalanced social groups and yield unfair representations. Critically, the unfairness cannot be trivially measured by losses. Thus, we propose a novel fairness metric based on feature balance to reveal the overlooked unfairness in pre-training.

Unsupervised fair representation learning. Different from its supervised version (Zemel et al., 2013), unsupervised fair representation learning aims to learn a fair feature space before supervision is available. Dutta et al. (2020) showed a surprising result that contradicts common perception of fairness in literature: given an ideal space, there could be a win-win of fairness and accuracy. Thus, they propose activate feature collection, which however could be expansive at practice. Park et al. (2020) leveraged protected and unprotected attributes together to learn representations disentangled from protected attribute. A similar work is Hwang et al. (2020). Ramapuram et al. (2021) first showed that self-supervised contrastive learning can benefit downstream fairness. Baldini et al. (2021) showed varying fairness of language models in downstream tasks and emphasized the importance of fairness mitigation during fine-tuning. Shen et al. (2021) studied the CL for fair representations when task attributes are available. Despite these primitive results, so far there is no study on how to measure and mitigate group unfairness without labels. The challenges are that we can hardly observe group disparity on in-group discrimination, nor retrieve extra labels in pre-training for group-fair learning. Our work addressed the two problems by a novel feature-based fairness metric and a novel fair learning algorithm.

2.1 Problem Setting and Preliminaries

This section elaborates the problem setting of interest. Let $\mathcal{X} = \{(x, a)\}$ denote an unlabeled source dataset and $\mathcal{D} = \{(x, y, a)\}$ the labeled target dataset, where x is a sample, y is its label, and a is its protected attribute. We use capital letters to denote random variables.

Unsupervised Pre-training + Supervised Fair Linear Fine-tuning. We study the powerful combination of unsupervised pre-training and a lightweight linear fine-tuning (PT+LFT). We will propose a fairness-aware solution that secures a fair feature representation from the pre-training phase, and uses it with a lightweight fair fine-tuning to generate fair downstream predictions.

$$\text{Pre-training: } \min_{\theta} \mathbb{E}_{(x,a)\in\mathcal{X}} \ell_{\text{CL}}(f_\theta(x), a), \tag{1}$$

$$\text{Fine-tuning: } \min_{\theta_c} \mathbb{E}_{(x,y,a)\in\mathcal{D}} \ell_{\text{CE}}^{\text{fair}}(\phi_{\theta_c} \circ f_\theta(x), y, a), \tag{2}$$

where ℓ_{CL} is a properly-designed CL loss aware of fairness and generalization, a function of the sample x, protected attribute a and feature encoder parameter f_θ, ϕ_{θ_c} denotes the classifier by equipping the linear prediction head ϕ_{θ_c}, $\ell_{\text{CE}}^{\text{fair}}$ is a fairness-aware cross-entropy loss. To ensure the efficiency of LFT, we use a simple fair scheme of group reweighing (Kamiran & Calders, 2012).

Contrastive learning. We adopt a representative contrastive setting SimCLR (Chen et al., 2020a). Given a batch of samples $\{x_1, \ldots, x_b\}$, features z are extracted by f_θ from two augmented views T_1

and T_2 as $\{f_\theta(T_1(x_1)), f_\theta(T_2(x_2)), \ldots, f_\theta(T_1(x_b)), f_\theta(T_2(x_b))\}$. The contrastive loss is

$$\ell_{\text{CL}}(z_i) = \sum\nolimits_{z_j \sim \mathcal{P}(i)} -\log \frac{s^\tau(z_i, z_j)}{\sum_{k \in \mathcal{N}(i)} s^\tau(z_i, z_k)}, \tag{3}$$

where $z_i = g \circ f(T_i(x))$ is the projected feature under the ith view; similarity function $s^\tau(a, b) = \exp(\cos\langle a, b\rangle / t)$; z_j is from the positive set; $\mathcal{N}(i)$ is the set of all views of all images except z_i, $2b - 1$ in total; $\mathcal{P}(i)$ is the positive set which includes different view of the ith sample.

Protect and assess downstream fairness. Regardless of the fairness in representations, downstream training could still ignore unprivileged groups because samples are under-presented. Therefore, we evaluate fairness after fair linear fine-tuning (FLFT). To protect fairness, we conduct a simple reweighing strategy (Kamiran & Calders, 2012):

$$\ell_{\text{CE}}^{\text{fair}} = P(A = a)/P(A = a|Y = y)\ell_{\text{CE}}, \tag{4}$$

where the variable probability can be estimated by statistic frequency. A very marginal complexity is required by reweighing to compute the group statistics in Eq. (4). Also, reweighing is easy to implement and does not trade-off hyper-parameters to tune. Therefore, reweighing provide a simple and efficient evaluation protocol. There are a variety of existing fairness metrics, and hereby we use a representative metric Δ_{EO} (Equalized Odds) as the fairness metric, whose disparity is defined as:

$$\Delta_{\text{EO}} \triangleq \sum\nolimits_{y \in \{0,1\}} \left| P(\hat{Y} \neq Y|A = 0, Y = y) \; - P(\hat{Y} \neq Y|A = 1, Y = y) \right|, \tag{5}$$

where $\hat{Y}$ is the predicted label, Y is the true label, and A denotes the protected attribute.

Cross-task fairness transferability. CL provides a way to learn representations from massive unlabeled datasets, which can be later utilized by a set of different downstream learning tasks through the fine-tuning process. The natural discrepancies between the two stages lead us to investigate the following challenging yet interesting quesitons: (1) Will CL yield unfair representations for downstream tasks? (2) If so, how to measure the inherent unfairness rooting from task-agnostic representation learning? (3) How can we transfer fairness from CL to different downstream tasks?

3 Challenges, Metric, and Mitigation

In this section, we answer the aforementioned questions and organize as follows: the challenges in fair contrastive learning, fairness metric without labels, and finally a mitigation method via minimizing the measured disparity. We illustrate the two major solutions in Fig. 2 where we introduce a fairness metric and a feature-centered mitigation method.

3.1 Challenge: The false sense of CL fairness

Though some evidence showed that CL improves downstream fairness (Ramapuram et al., 2021; Pruksachatkun et al., 2021), it is unclear if CL is immune to unfairness, and its fairness impact on downstream tasks with commonly-used linear fine-tuning.

Experiment setup for measuring representation fairness. Based on the PT+LFT paradigm in Section 2.1, we investigate the fairness impact on downstream classifiers from representations, or *representation fairness*. For fair representations, the feature vectors from both groups are similarly separable or of the similar best accuracy. As such, fair representations may correspond to an ideal construct space enabling the jointly achievable accuracy and fairness (Dutta et al., 2020). However, the data or label distribution in downstream tasks can also impact the fairness during the LTF procedure, and therefore we consider fair LFT instead of standard LFT, such that we can measure the desired representation unfairness instead of algorithmic unfairness.

Matched contrastive loss does not imply fair representations. We note a critical challenge here is that a straightforward investigation of the CL loss may create an illusion that CL is always fair. In Fig. 1, we show that CL loss distributes similarly between two groups, which suggests the the representation *should be* fair according to the conventional wisdom from existing studies. However, our study demonstrated a surprising result that group imbalance in CL pretraining result in downstream unfairness, even when existing fairness measures over CL loss suggested otherwise. We conduct experiments on the CelebA dataset and measure the group-balanced accuracy (B-Accuracy), i.e., the average of group accuracy, and fairness Δ_{EO}, under different group ratios (male/female).

In Fig. 3, we show that fairness degrades when the group ratio in training CL gets increasingly imbalanced, while accuracy remains similar.

3.2 METRIC: FEATURE DISPARITY

Recall the goal of CL is to learn features versatile for multiple downstream supervised tasks. We conjecture that, the reason that a group performs worse than the other in a downstream task, is because the group learns fewer discriminative features, and therefore the downstream supervised learning cannot find a capable linear decision boundary on such feature space. For example, in a learning task the male group may have fewer samples of long hairs and CL ignores the hair features, resulting in poor recognition of the length of hairs. Consistent with the conjecture, an existing theoretical result has shown that collecting more task-dependent features for the under-presented group is in favor of the fairness or fairness-accuracy trade-off (Theorem 3 in (Dutta et al., 2020)).

Therefore, we believe that the feature sufficiency could be closely related to the downstream performance and therefore task-specific fairness. Without the awareness of tasks, we characterize the unfairness by a novel principle:

$$\text{Unfairness} \triangleq \Delta(\#\text{Mastered Features}), \tag{6}$$

where $\Delta(n)$ describe the group difference of n, and a feature is *mastered* if it contributes significantly to lowering the loss values, and therefore is critical in the sense of self-supervised discrimination. Let $M \in [0,1]^C$ be a mask vector where C is the dimension of a feature vector z. We define the feature influence by the approximated loss perturbation on removal:

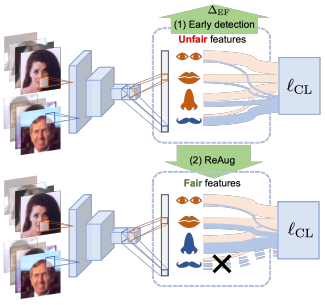

Figure 2: Illustration of unfair features in male/female groups and the mitigation. (1) The potential downstream classification unfairness can be exposed by the imbalanced influential features in female and male groups. (2) Re-augment unprivileged group features and arrive at fair representations.

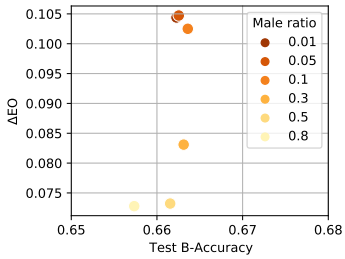

Figure 3: Unfairness induced by imbalanced pre-training. Results are averaged over multiple tasks of CelebA.

$$I^i(z) \triangleq \left|\ell(\mathbf{1}_{\setminus i} \circ z) - \ell(\mathbf{1} \circ z)\right| \approx \left|\frac{\partial \ell(M \circ z)}{\partial M^i}(0 - M_i)\right|_{M=\mathbf{1}} = \left|z_i \frac{\partial \ell(z)}{\partial z_i}\right|, \tag{7}$$

where $\mathbf{1}_{\setminus i}$ is all-one vector except dimension i. Eq. (7) has two factors: (1) *discrimination:* The gradient reflects the influence of the feature on local discrimination by the definition of the contrastive loss. (2) *activation:* numerical scale of the feature indicates the activation influence. If an image is not discriminative from others, for example, all females do not have beard, then the feature will not activated. Instead, if all male and most female activate the big eye feature, then the feature will be activated but not discriminative. In both cases, the feature is less important and of quantitively minor I^i. When a feature is always important in a data set, then feature is significantly learned. Therefore, for a group $a \in \{0,1\}$ with a distribution P, we define its *group feature influence* by:

$$\mathcal{I}^i(P_{A=a}) = \mathbb{E}_{x \sim P(x|A=a)}[I^i(z)].$$

With the definition of group feature influence, we define a group feature to be ϵ-*influential* if $\mathcal{I}^i(P_{A=a}) > \epsilon$ given a threshold ϵ. Then we empirically show the feature disparity in Fig. 4. The under-presented group (male) has 98% fewer samples and learns 24% (122) fewer influential features than the dominant group.

To quantify Eq. (6) without parameterization of ϵ, we consider the maximal disparity of the influential-feature ratios between groups and therefore conclude our metric in Definition 3.1 following the principle of Eq. (6).

Definition 3.1 (Feature Disparity)**.** *A self-supervised representation is considered to be fair if the maximal disparity of features, as computed below, is small:*

$$\Delta_{\text{FD}} = \max_{\epsilon} \left|\mathbb{P}\left[\mathcal{I}(P_{A=0}) > \epsilon\right] - \mathbb{P}\left[\mathcal{I}(P_{A=1}) > \epsilon\right]\right|,$$

where $\mathbb{P}(x) = \frac{1}{C}\sum_{i=1}^{C} \mathbb{I}(x^i)$*. Meanwhile, a group* 0 *is considered to be* unprivileged *if*

$$\mathbb{P}\left[\mathcal{I}(P_{A=0}) > \epsilon\right] < \mathbb{P}\left[\mathcal{I}(P_{A=1}) > \epsilon\right].$$

The definition can be easily extended to multiple group cases by using the most disparate groups, i.e.,

$$\max_{\epsilon} \left\{ \max_{a} \mathbb{P}\left[\mathcal{I}(P_{A=a}) > \epsilon\right] - \min_{a} \mathbb{P}\left[\mathcal{I}(P_{A=a}) > \epsilon\right] \right\}.$$

The complexity of maximization in Δ_{FD} is $\mathcal{O}(C \log C)$ which can be done by sorting the influences and traversing ϵ over all $\mathcal{I}^i$.

To clarify the precautionary unfairness related with the proposed Equalized Feature, we provide the following remark.

Remark 3.1. *If Δ_{FD} is sufficiently greater than zero, then there exist downstream tasks which are unfair in the notion of Δ_{EO}.*

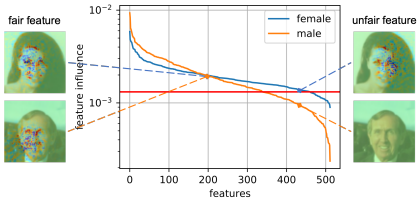

Figure 4: Feature influence differs by group on imbalance CelebA with a SimCLR-trained model. The x-axis represents the indexes of features sorted according by their influence. The red line shows the threshold maximizing the feature disparity, above which the female group has more significant features than the male group. Features are attributed to the original images by GradCAM when the target channel is kept and others are masked out (Selvaraju et al., 2016).

The remark is based on the intuition: every downstream task is dependent on a subset of the learned features and a non-zero Δ_{FD} implies that a group may lacks features in the subset. On the other hand, it is worth to mention that $\Delta_{\text{FD}} = 0$ does *not* imply downstream fairness. A counter-fact is that when each group has mastered the same number of features yet without overlapping, then Δ_{FD} is zero but unfair in downstream tasks.

3.3 Method: Re-augment features (ReAug)

As Δ_{FD} in Definition 3.1 is non-differentiable, we provide a practical method to reduce Δ_{FD}. The major idea is to encourage the unprivileged group to learn more features to mitigate the disparity.

Feature augmentation by masked training. Masked training has been developed for improving the generalization of supervised (Huang et al., 2020) or self-supervised learning (Mo et al., 2021). Both methods mask out high-influential features iteratively for each sample and minimize losses on the rest features. However, they cannot be directly applied to optimize the proposed metric due to the group nature of Δ_{FD}. To this end, we propose a group-wise masking strategy to optimize the loss over masked features, following the proposed equalized feature principle:

$$L_{\text{mask}} = \mathbb{E}_{x,A}[\mathbb{I}(A=0)\ell_{\text{CL}}(M_{A=0} \circ f_\theta(x)) + \mathbb{I}(A=1)\ell_{\text{CL}}(M_{A=1} \circ f_\theta(x))], \tag{8}$$

where $M \in [0,1]^C$. To see how the strategy works, let us suppose group 1 is the unprivileged one, suggesting that group 1 mastered $\Delta_{\text{FD}} \times 100\%$-fewer features by Definition 3.1. In order to protect group 1, the strategy drops its most influential $\Delta_{\text{FD}} \times 100\%$ features and maintains all features of the other group:

$$M^i_{A=1} = \mathbb{I}\left[\mathcal{I}^i_{A=1} < \mathcal{I}^q_{A=1}\right],\ M^i_{A=0} = 1, \tag{9}$$

where $\mathcal{I}^q_{A=1}$ is the largest $\mathcal{I}^i_{A=1}$ such that $\mathbb{P}[\mathcal{I}^i_{A=1} < \mathcal{I}^q_{A=1}] \leq \Delta_{\text{FD}}$ and $q = \Delta_{\text{FD}}$.

One issue of this group-wise masking strategy is that it may hardly learn new features for minority groups. It is easy to see that group-wise masking increases the dissimilarity of two different samples, i.e., $s^\tau(M^i_{A=0} \circ z_i, M^i_{A=1} \circ z_k) < s^\tau(M^i_{A=0} \circ z_i, M^i_{A=1} \circ z_i)$ if $(1 - M^i_{A=1}) \circ z_i > 0$. As a result, the retained features will not learn new discrimination knowledge by contrasting inter-group samples but intra-group samples. But the sample insufficiency of the minority group makes it difficult to develop new features.

Global masking to enhance feature transfer. To introduce new information for the minority, we propose a global masking: setting $M^i_{A=0}$ equal $M^i_{A=1}$. First, by masking most of the minority's interest, we encourage the minority samples to learn more discriminative features not only from intra- but also inter-group negative samples. Second, the majority group will also pay attention to discriminate samples by the selected features and therefore the influence of these features will be augmented. Last, a byproduct of the global masking is the removal of the potentially over-learned inter-group discrimination. In a contrastive loss, a minority-group sample will naturally have more contrastive samples from the majority group than the minority one. Due to the scarcity of the minority

samples and a variety of different features between the two groups, the discriminative learning will quickly learn few influential features that simplify the functional for discriminating groups. Therefore, the global masking biased toward the minority helps the group to get rid of such inter-group features.

Efficient group feature influence estimation. Estimating feature influence requires going through all samples and computing gradient (at least a forwarding-time complexity), which doubles the training time. To reduce the overhead, we propose to use momentum to estimate the feature influence, which is almost as efficient as standard CL:

$$\hat{\mathcal{I}}^i_{t+1} = (1-\beta)\hat{\mathcal{I}}^i_t + \beta \mathbb{E}_{x\sim\mathcal{B}}[I^i(z)], \tag{10}$$

where β is the momentum parameter (0.1 if not specified).

Delayed feature-masking. The feature masking relies on the assumption that a group has a stable influence. Crafting masking at random initialization or an early learning stage could lead to unwanted masking or random masking, which prevents the training to learn features. Therefore, we first train an encoder with SimCLR (Chen et al., 2020a) and fine-tune the features with masking.

In summary, our proposed method includes a new masked feature loss and a three-stage training algorithm, as described in Algorithm 1. The overall complexity of masking and updating parameters, e.g., by gradient descent, is $\mathcal{O}(dBC^2 + C\log C)$. The major component of complexity is gradient computation, approximately as $\mathcal{O}(dBC^2)$ given batch size B, feature size C and a d-layer equal-width MLP projection head (for computing CL loss). Thanks to the shallow structure of the contrastive projection head, the gradient is not too expensive. The second term in the complexity is of an efficient implementation, e.g., of searching for the maximal ϵ in the feature influence array.

Algorithm 1 Three-stage Training

1: **Input:** Unlabeled dataset and few-labeled dataset
2: Stage 1: Train f_θ by SimCLR.
3: Stage 2: Iteratively update feature influence by Eq. (10) and train f_θ by masked CL loss Eq. (8).
4: Stage 3: Fairly fine-tune linear classifier over frozen f_θ.
5: **Output:** A fair classifier and a fair encoder.

4 Empirical Results

In this section, we conduct extensive experiments to demonstrate the effectiveness of the proposed method and metric from two folds. **(1) Quantitative studies** include cross-task fairness transferability, cross-dataset fairness transferability, and fairness against different levels of group imbalance. **(2) Qualitative studies** includes the analysis of contrastive loss distributions to understand the effects of ReAug and how the proposed unsupervised fairness metric Δ_{FD} serves as a predictive metric for downstream fairness.

Experiment setup. We evaluate the representation fairness through two *metrics* over fair linear fine-tuning (FLFT): (1) Fairness metric Δ_{EO} where a smaller value indicates a fairer model. (2) Balanced accuracy (B-Acc) computes the mean of group-wise accuracy which avoids ignoring a minority group at test time. We compare our method with a standard CL algorithm, SimCLR. We evaluate methods on multiple face datasets with sensitive attributes: CelebA (Liu et al., 2015), UTK (Zhang et al., 2017), FairFace (Karkkainen & Joo, 2021). We use CelebA for pre-training and conduct FLFT evaluation on labeled CelebA, FairFace, and UTKFace predicting binary attributes. As stated in Section 2.1, we use the simple and efficient reweighing (Kamiran & Calders, 2012) as the unfairness mitigation. LAFTR (Madras et al., 2018) is also used in the binary prediction task on CelebA.

Training. During pre-training, we train a ResNet18 model (He et al., 2016) by SimCLR (Chen et al., 2020a) for 1000 epochs and continue to train for extra 1000 epochs with or without ReAug. Following the common practice, we use a two-layer equal-width projection network upon the representations, before computing the CL loss. Per iteration, we use a batch of 1024 samples to compute the gradient, and the learning rate is decreased by cosine-annealing from 10^{-3} to zero. With the gradient, Adam (Kingma & Ba, 2015) with 10^{-5} weight decay is utilized for both optimizing contrastive losses and linear fine-tuning. During fine-tuning, we freeze feature extractor layers (e.g., all convolutional layers of ResNet), and fine-tune the last linear layer for 30 epochs with a batch of size 128 and the same learning rate schedule as pre-training. The trade-off parameter of LAFTR is set as 5 for more fairness.

Table 1: CelebA FLFT cross-task evaluation with 1% and 100% FLFT data with reweighing and LAFTR mitigation.

Mitigation	Reweighting				LAFTR			
	$\Delta_{\text{EO}}(\%)\downarrow$		B-Acc $(\%)\uparrow$		$\Delta_{\text{EO}}(\%)\downarrow$		B-Acc $(\%)\uparrow$	
Task	SimCLR	+ReAug	SimCLR	+ReAug	SimCLR	+ReAug	SimCLR	+ReAug
				1% FLFT				
Attractive	15.5	**14.1**(-1.4)	73.5	**73.9**(+0.5)	80.5	**77.4**(-3.1)	75.3	**76.3**(+1.0)
Big Lips	**9.5**	10.6(+1.0)	53.4	53.4(+0.0)	23.4	**8.1**(-15.3)	**85.4**	85.2(-0.2)
Black Hair	9.3	**7.7**(-1.6)	**75.4**	75.1(-0.3)	**5.1**	6.3(+1.2)	**86.7**	85.9(-0.8)
Brown Hair	13.8	**11.6**(-2.2)	**56.8**	56.4(-0.4)	**30.4**	38.6(+8.2)	**81.7**	81.6(-0.1)
High Cheekbones	12.3	**11.0**(-1.3)	82.2	**82.4**(+0.2)	21.7	**18.2**(-3.5)	**84.4**	83.5(-0.9)
Mouth Slightly Open	7.7	**7.2**(-0.5)	77.7	**78.5**(+0.7)	13.9	**10.2**(-3.7)	**78.6**	78.4(-0.2)
Narrow Eyes	0.2	0.2(+0.0)	50.1	50.1(-0.0)	**2.2**	2.9(+0.7)	92.4	92.4(+0.0)
Oval Face	14.9	**10.1**(-4.9)	**56.0**	55.6(-0.4)	33.0	**31.5**(-1.5)	73.5	**73.8**(+0.3)
Pointy Nose	2.3	**1.5**(-0.8)	**55.9**	55.5(-0.4)	35.2	**32.5**(+0.3)	73.1	**73.4**(+0.3)
Smiling	7.7	**6.8**(-0.9)	88.3	**89.1**(+0.8)	12.6	**7.2**(-5.4)	88.1	88.1(+0.0)
Straight Hair	7.3	**5.8**(-1.5)	**54.2**	53.2(-1.1)	**9.2**	14.1(+4.9)	80.8	**81.2**(+0.4)
Wavy Hair	23.9	**20.9**(-3.0)	**69.8**	69.4(-0.4)	**49.2**	**57.5**(+8.3)	81.1	**82.5**(+1.4)
Young	**5.4**	8.5(+3.0)	62.5	**63.5**(+0.9)	42.2	**34.0**(-8.2)	**83.0**	82.2(-0.8)
Average	10.0	**8.9**(-1.1)	65.8	65.8(+0.0)	27.6	**26.0**(-1.6)	81.9	81.9(+0.0)
				100% FLFT				
Attractive	8.6	**7.7**(-0.9)	75.6	75.6(-0.1)	79.0	**76.2**(-2.8)	75.3	**76.6**(+1.3)
Big Lips	**17.9**	18.5(+0.6)	54.2	54.2(+0.0)	23.3	**6.3**(-17)	**85.3**	85.1(-0.2)
Black Hair	7.9	**6.0**(-1.9)	79.8	**80.8**(+1.0)	**4.0**	6.1(+2.1)	**86.8**	85.6(-1.2)
Brown Hair	22.2	**19.1**(-3.1)	72.3	**73.3**(+1.1)	**30.5**	39.3(+8.8)	**81.6**	81.5(-0.1)
High Cheekbones	12.4	**10.9**(-1.5)	83.4	**83.8**(+0.3)	21.4	**19.1**(-2.3)	**84.4**	83.4(-1.0)
Mouth Slightly Open	8.6	**7.5**(-1.1)	79.6	**79.9**(+0.3)	14.1	**10.5**(-3.6)	**78.8**	78.6(-0.2)
Narrow Eyes	**4.3**	4.7(+0.4)	51.6	**52.0**(+0.4)	**2.0**	3.0(+1.05)	92.4	92.4(+0.0)
Oval Face	**15.6**	17.1(+1.6)	57.6	**57.9**(+0.3)	34.9	**31.4**(-3.5)	73.5	**73.8**(+0.3)
Pointy Nose	4.0	**2.8**(-1.1)	**58.5**	58.2(-0.3)	36.1	**32.9**(-3.2)	73.2	**73.5**(-0.3)
Smiling	7.4	**6.3**(-1.1)	90.0	**90.2**(+0.2)	13.0	**7.7**(-5.3)	88.1	88.1(+0.0)
Straight Hair	**20.3**	20.9(+0.6)	60.9	60.9(+0.0)	**8.7**	14.0(+5.3)	80.8	**81.2**(+0.4)
Wavy Hair	23.9	**16.8**(-7.1)	75.4	**75.6**(+0.2)	**49.8**	57.4(+7.6)	80.7	**82.4**(+1.7)
Young	1.7	**1.4**(-0.2)	67.2	**68.4**(+1.2)	41.0	**33.2**(-7.8)	**82.9**	82.2(-0.7)
Average	11.9	**10.7**(-1.1)	69.7	**70.1**(+0.4)	27.5	**25.9**(-1.6)	81.8	**81.9**(+0.1)

4.1 Quantitative Results

In this subsection, we compare the fairness of representations in two scenarios: fine-tune the model on labeled pre-training dataset (cross-task), and on labeled unseen dataset (cross-dataset). The former evaluates how well the fairness can transfer from the unsupervised task to different supervised ones. The latter considers a general case when a pre-trained large model is used for fast adaptation to a target dataset. In addition, we re-sample the CelebA dataset to have 1% Male and 99% female to exacerbate the unfairness risk in pre-training and thus we can evaluate pre-training algorithms in a more challenging setting. Such an extreme imbalance is in fact rather common in practice. For example, most states in the United States have less than 1% Non-Hispanic American Indian, according to U.S. Census in 2017. Therefore when using medical records to study diseases and treatments, we are very likely to encounter extreme imbalance. Then the FLFT evaluation protocol is conducted on a 1% balanced downstream dataset where we resample the dataset to make the groups are balanced, and a 100% naturally imbalanced same dataset. The few-sample setting is commonly considered for adaptation to data-insufficient domains. For example, a home-located sensor may not have enough data during early deployment.

Cross-task results. In Table 1, we present the cross-task evaluation results on CelebA, where 13 tasks are selected to enclose enough samples for testing (if not so, we may get biased evaluation). When only 1% data are available for fine-tuning, the proposed learning method improves SimCLR in 10 out of 13 tasks, when balanced accuracy (B-Accuracy) is not explicitly degraded. The improvement is more significant with severer downstream unfairness. For example, the top-two improvement happens in Oval Face and Wavy Hair tasks, where +ReAug improves SimCLR by 4.9% and 3% Δ_{EO}, respectively. As more data are available for FLFT, the fairness gain from ReAug is less significant than in the former case but ReAug improves the accuracy in more cases.

Table 2: CelebA cross-dataset evaluation with 1% or 100% FLFT data for FLFT. We use gender as the sensitive attribute.

Data	Task	$\Delta_{\text{EO}}(\%)\downarrow$ SimCLR	$\Delta_{\text{EO}}(\%)\downarrow$ +ReAug	B-Acc $(\%)\uparrow$ SimCLR	B-Acc $(\%)\uparrow$ +ReAug
			1% FLFT		
FairFace	Young	26.4	**25.7**(-0.6)	68.3	**69.3**(+1.0)
UTKFace	Young	17.8	**12.0**(-5.8)	48.4	**49.1**(+0.7)
			100% FLFT		
FairFace	Young	33.9	**26.2**(-7.7)	71.1	**72.5**(+1.4)
UTKFace	Young	19.2	**11.9**(-7.3)	48.3	48.2(-0.0)

Different unfairness mitigation approaches have different influences on performance improvement. With 1 % LAFTR mitigation, Big Lips and Smiling have the most distinguish improvement, where Δ_{EO} has been improved by 15.3% and 5.4%, respectively. While there is less improvement in accuracy under LAFTR, fairness improvement is enhanced compared with reweighing mitigation.

Cross-dataset results. Table 2 presents the results of fairness transfer from CelebA dataset to FairFace or UTKFace datasets. We use CelebA as the pre-training dataset which does not utilize labels. We fine-tune the model on a small subset (1%) or full set (100%) of FLFT datasets to evaluate the fairness of the representations. Gender is used as the sensitive attribute and Young (for ages lower than 30) is the task attribute. In all FLFT datasets, our method consistently improves classification fairness. The largest improvement occurs with the FairFace by $7.7\%\Delta_{\text{EO}}$ when accuracy is improved, as well.

4.2 Qualitative Results

We provide qualitative studies to understand the proposed method. All results are done with pre-training on CelebA and fine-tuning on 1% labeled CelebA dataset including previously discussed results in Figs. 1 and 3.

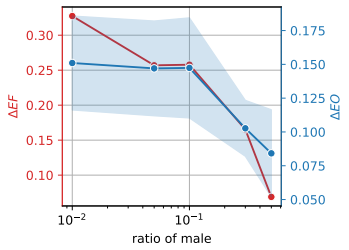

Figure 5: The pre-training (Δ_{FD}) and fine-tuning (Δ_{EO}) disparity versus the ratio of male in pre-training. A smaller ratio of male indicates higher degree of imbalance.

ReAug masking implicitly upweighs minority losses. We are interested in seeing how the proposed ReAug method affects the sample losses in CL. Instead of using $q = \Delta_{\text{FD}}$ in Eq. (9), we temporarily consider q as a variable that serves as the drop rate of features. We define the loss disparity as:

$$\{\mathbb{E}_x[L(x)|A=1] - \mathbb{E}_x[L(x)|A=0]\} / \mathbb{E}_x[L(x)],$$

where group 1 is assumed to be the unprivileged group (male in our experiments). The result is depicted in Fig. 6. Though not intently designed, we find that the proposed ReAug method implicitly changes the losses from different groups and augments the disparity of losses between minority and majority groups, when masking out more features. In the beginning, because the masking removes the most important male features, losses for the male group increase faster than that for the female group. When too many features are masked out, the loss disparity tends to vanish w.r.t. q, because fewer influential features are maintained. Note that our method is not a trivial reweighing solution, and the masking strategy resembles the dropout to help generalization. On the other hand, an explicit reweighing method would encourage the model to overfit the minority group and therefore would not generalize, especially when the minority group is trained as well as the majority group in CL.

Δ_{FD} predicts the unfairness in downstream. To examine the relation between CL Δ_{FD} and downstream Δ_{EO}, we select tasks that have severe fairness degradation (Δ_{EO} is larger than 0.1 after imbalance pre-training) and show how the degradation is related to pre-training data and features. In Fig. 5, task-averaged Δ_{EO} is plotted versus the ratio of minority group varying from 1% (very unbalanced) to 50% (balanced). Consistent with our intuition, disparity both in features (Δ_{FD}) and downstream tasks (Δ_{EO}) decreases as the imbalance degree vanishes. The result motivates us to exploit feature disparity in pre-training in favor of downstream fairness.

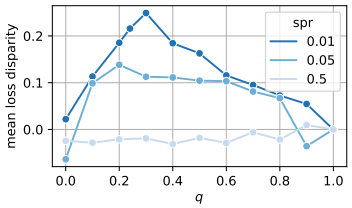

Figure 6: CL loss disparity versus feature drop rate q. Masking implicitly augments the overlooked disparity in losses. `spr` denotes the sensitive positive rate whose small value indicates an imbalance dataset.

5 Conclusion and Discussions

In this paper, we first showed the underestimated unfairness caused by group imbalance in the pre-training phase, and proposed an unsupervised fairness metric to detect such bias early before supervisions are available. To our best knowledge, our work is the first attempt to measure and address precautious fairness in pre-training. With the growing importance of large pre-training models and the rising concerns of fairness, our work provides a practical solution that can benefit multiple application areas. One future direction is to extend the feature-based metric and mitigation to the natural language processing area in which large pre-training models are in demand.

REFERENCES

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. Your fairness may vary: Group fairness of pretrained language models in toxic text classification. *arXiv:2108.01250 [cs]*, August 2021.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18, December 2009.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, pp. 13, 2020b.

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

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297 [cs]*, March 2020c.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *arXiv:1802.08626 [cs, stat]*, January 2020.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, December 2020.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6391–6400, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv:2006.07733 [cs, stat]*, September 2020.

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, pp. 3323–3331, Red Hook, NY, USA, December 2016. Curran Associates Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition*, 2016.

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*, pp. 9729–9738, 2020.

J. Henry Hinnefeld, Peter Cooman, Nat Mammo, and Rupert Deese. Evaluating fairness metrics in the presence of dataset bias. *arXiv:1809.09245 [cs, stat]*, September 2018.

Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *ECCV*, July 2020.

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*, 2020.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012.

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

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1547–1557, Waikoloa, HI, USA, January 2021. IEEE.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 247–254, New York, NY, USA, January 2019. Association for Computing Machinery.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *the 3rd International Conference for Learning Representations*, San Diego, CA, 2015.

Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2737–2746. PMLR, July 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pp. 502–510, New York, NY, USA, August 2011. Association for Computing Machinery.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3384–3393. PMLR, July 2018.

Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *arXiv:2108.00049 [cs]*, October 2021.

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

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3320–3331, Online, August 2021. Association for Computational Linguistics.

Jason Ramapuram, Dan Busbridge, and Russ Webb. Evaluating the fairness of fine-tuning strategies in self-supervised learning. October 2021.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, September 2002.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9419–9427, May 2021.

Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. Transfer of machine learning fairness across domains. June 2019.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv:1610.02391 [cs]*, October 2016.

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

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 3–13, New York, NY, USA, March 2021. Association for Computing Machinery.

Dylan Slack, Sorelle A. Friedler, and Emile Givental. Fairness warnings and fair-maml: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 200–209, New York, NY, USA, January 2020. Association for Computing Machinery.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv:2005.10243 [cs]*, December 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748 [cs, stat]*, January 2019.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. *CVPR*, March 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939. PMLR, November 2020.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 1920–1953. PMLR, June 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1171–1180, Republic and Canton of Geneva, CHE, April 2017a. International World Wide Web Conferences Steering Committee.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 962–970. PMLR, April 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333. PMLR, May 2013.

Tao Zhang, tianqing zhu, Jing Li, Mengde Han, Wanlei Zhou, and Philip Yu. Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

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*, pp. 5810–5818, 2017.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13208–13217, 2020.