Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University

Abstract

Recently, fairness-aware learning have become increasingly crucial, but most of those methods operate by assuming the availability of fully annotated demographic group labels. We emphasize that such assumption is unrealistic for real-world applications since group label annotations are expensive and can conflict with privacy issues. In this paper, we consider a more practical scenario, dubbed as Algorithmic Group ***Fair****ness with the* ***P****artially annotated* ***G****roup labels (**Fair-PG**). We observe that the existing methods to achieve group fairness perform even worse than the vanilla training, which simply uses full data only with target labels, under Fair-PG. To address this problem, we propose a simple* ***C****onfidence-based* ***G****roup* ***L****abel assignment (**CGL**) strategy that is readily applicable to any fairness-aware learning method. CGL utilizes an auxiliary group classifier to assign pseudo group labels, where random labels are assigned to low confident samples. We first theoretically show that our method design is better than the vanilla pseudo-labeling strategy in terms of fairness criteria. Then, we empirically show on several benchmark datasets that by combining CGL and the state-of-the-art fairness-aware in-processing methods, the target accuracies and the fairness metrics can be jointly improved compared to the baselines. Furthermore, we convincingly show that CGL enables to naturally augment the given group-labeled dataset with external target label-only datasets so that both accuracy and fairness can be improved. Code is available at* https://github.com/naver-ai/cgl_fairness.

1. Introduction

Recent advances of machine learning (ML) models have witnessed promising outcomes even in societal applications, such as credit estimation [28], crime assessment systems [7, 23], automatic job interviews [33], face recognition [8, 38], and law enforcement [17]. However, machines are often more inaccurate to a particular group (*e.g.*, darker-skinned females) than other groups (*e.g.*, lighter-skinned males) [8], *i.e.*, machines are discriminatory. To mitigate the issue, *fairness-aware learning* has recently emerged; a model should not discriminate against any demographic group with sensitive attributes, *e.g.*, age, gender, or race.

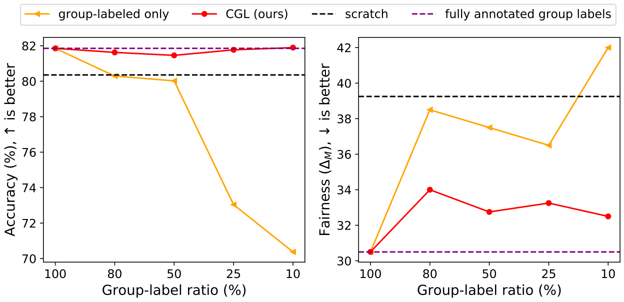


Figure 1. **Can fair-training methods still learn fair classifiers when group labels are partially annotated?** We note the state-of-the-art fairness fair-training FairHSIC [34] using only the group-labeled subset (yellow) shows worse fairness criterion (Δ_M, Eq. (2), lower the better) than the "scratch" (*i.e.*, no consideration of a fairness criteria) in the low group label regime (*e.g.*, 10%) on UTKFace [44]. Our CGL (red), on the other hand, can be potentially applied to any fair-training method, and when it is combined with FairHSIC, both the target accuracy and the fairness criteria are significantly improved for the low group label regime.

Many existing approaches for *group fairness* [1, 12, 22, 24, 34, 41, 42] utilize two types of labels: *target labels*, which are task-oriented (*e.g.*, crime assessment) and *group labels*, which are defined by socially sensitive attribute groups (*e.g.*, ethnicity or gender). Many existing methods for achieving group fairness rely on the group labels to train fair classifiers. For example, many approaches explicitly minimize the statistical parity metrics between groups defined by sensitive attributes. However, in many realistic applications, *e.g.,* computer vision, assuming that all images have sensitive group labels can be unrealistic and make the existing methods impractical. First, in many image datasets, group labels are not explicitly given as in tabular datasets [7, 23, 28] but are defined in high-level se-

[*] Works done while doing an internship at NAVER AI Lab.
[†] Corresponding authors

mantics, requiring additional expensive human annotations. Secondly, the sensitive attributes are usually personal information protected by laws, such as EU General Data Protection Regulation (GDPR). Hence, in real-world applications, collecting group labels for all data points are impossible without permissions by all users and, furthermore, sensitive attributes cannot be persistently stored but should be expired. Thus, the underlying assumption by the previous fair-training methods, *i.e*., group labels are fully annotated, can limit their usability in real-world applications.

Contribution. In this work, we propose and investigate a less explored but very practical problem: *Algorithmic Group **Fair**ness with the **P**artially annotated **G**roup labels (**Fair-PG**)*. Many existing fair-training methods for group fairness assume all training samples have group labels, and optimize fairness constraints by the group labeled training samples. In this case, they cannot be directly applied to the Fair-PG problem. We empirically show that the baseline fair-training methods, which operate only on the group-labeled samples, perform even worse than the vanilla "scratch" training that use all the training samples, in terms of fairness when the number of group-labeled samples is small (*e.g*., 10%) – See Fig. 1. Although there exist a few attempts to achieve algorithmic fairness without demographics labels [20, 29], they do not directly solve the *group fairness* problem. Also, they do not utilize partially annotated group labels at all, while a small number of labeled data can improve the overall performances. To this end, we propose a simple yet effective strategy for Fair-PG that can be applied to *any* fair-training methods for group fairness, dubbed as **C**onfidence-based **G**roup **L**abel assignment (**CGL**). CGL assigns pseudo group labels to group-unlabeled samples using an auxiliary group classifier, if the predictions are sufficiently confident, and random group labels, otherwise.

We provide high-level understandings of how CGL works on the Fair-PG scenario. We theoretically support that (1) the fairness parity computed by our approach approximates the parity of the underlying group label distribution better than the one by the vanilla pseudo-label strategy which totally trusts the predictions of the auxiliary group classifier, (2) assigning a random group label to a data point implies the elimination of the fairness constraint of the sample. In practice, since the existing fair-training methods use a relaxed constraint, CGL can be interpreted as a regularization method for the low confident group-unlabeled samples.

In our experiments, the combination of CGL with state-of-the-art fair-traning methods (*e.g*., MFD [24], FairHSIC [34] and LBC [22]) has consistently and significantly improved target accuracies as well as fairness parities even under the low group label regime on facial image [31, 44] and tabular [14, 23] datasets. For example, compared to the "group-labeled only" baseline, the combination of CGL and MFD shows +8.23% target accuracy increase and -8.75 disparity of equal opportunity (DEO) decrease on UTKFace [44], when only 10% of data points have group labels. Further extending this result, by augmenting the full UTKFace training set with extra group-unlabeled dataset in [27], we show that CGL can significantly improve the performance of MFD by +0.92% accuracy and -5.5 DEO. This is promising since it shows CGL can improve both the accuracy and fairness of a baseline method by augmenting the training data with target label-only dataset, which is relatively easier to obtain than jointly requiring the group labels.

2. Related Works

Fair-training for group fairness. There have been various works to tackle fairness problem in machine learning models. At a high level, it can usually be classified into three categories, including 1) *individual fairness* [15, 40] that aims to treat similar users similarly, 2) *group fairness* [19] of which goal is to reduce the statistical parity between groups defined by the sensitive attributes, 3) *Rawlsian minmax fairness* [16,20,29] which designs to improve the worst performance among groups. In this paper, we follow the notion of *group fairness* in arguing the fairness of a model. Many fair-training methods have been developed to achieve group fairness. The fairness methods (for group fairness) can be divided into three categories depending on where the technique for fairness is injected into; *pre-processing* methods [13, 34, 42] modify a training dataset before learning a model; *in-processing* [12, 22, 24, 26, 41, 43] methods consider fairness during training time; *post-processing* methods [3] modify a trained model. However, despite technical advances for achieving group fairness, existing methods for group fairness have not considered the setting in which a part of a training dataset lacks group label (*i.e*., Fair-PG).

Fairness with imperfect sensitive attributes. Recently, a few attempts have been proposed to consider algorithmic fairness with imperfect sensitive attributes, *e.g*., noisy group labels. Chen *et al*. [10] and Kallus *et al*. [25] proposed methods for assessing the disparity when only proxy variables (*e.g*., surname) for the protected variables are given. Meanwhile, several works [2, 39] posed solutions of learning a fair classifier robust to *noisy* group labels. However, these approaches focus only on *noisy* group labels, hence they cannot be directly applied to our Fair-PG setting.

There also have been a few works for fair-training without any information of protected attributes. Hashimoto *et al*. [20] proposed a distributionally robust optimization (DRO) [32] based approach, and Lahoti *et al*. [29] utilized an adversary to identify regions with high loss and re-weight them. Similarly, there exist *de-biasing* methods to solve the bias problem without any labels denoting bias (discussions given in the Appendix). However, these methods have two limitations compared to CGL on Fair-PG. First, they do not directly aim to achieve *group fairness*, but they consider the *Rawlsian Min-Max fairness* or *cross-bias generalization*. Second, it is not straightforward to combine them to the existing fair training methods, while our method can be universally applicable to any fair-training method.

Semi-supervised learning. Semi-supervised learning (SSL) [9] aims to learn a model with a small number of labeled samples and a large number of unlabeled samples. Our Fair-PG scenario also considers when group labels are partially annotated but target labels are fully annotated. However, since the aim for SSL is to simply *predict* the future attribute labels as accurately as possible from the partial annotations in the training set, it is not clear whether the predicted attribute labels can be directly plugged-in to achieve the *group fairness* in the test set. In addition, the recent state-of-the-art SSL methods [5, 6, 37] are hardly applicable to the fairness problem directly because they mostly focus on seeking better augmentation methods and consistent constraints for the augmented inputs. Instead, our method is motivated from the pseudo-labeling (PL) strategy [30] to avoid complex modifications on the base fair-training methods by assigning pseudo group labels to the group-unlabeled samples. While the original PL fully trusts the network predictions, we set the random labels for the low confident samples. We show, both theoretically and empirically, that such *random label selection strategy* in CGL is critical for achieving better group fairness, and we believe such finding is not straightforward. Our strategy is also similar to the recently proposed UPS [35] in the SSL context, which withdraws pseudo-labels for low confident samples by using an external uncertainty prediction module. However, CGL uses the full group-unlabeled samples by the random label strategy and outperforms the UPS-base strategy in our experiments.

3. Problem Definition

In this section, we formally define our scenario, Fair-PG, and the fairness criterion, *disparity of equality of opportunity (DEO)*, for the general M-ary classification problem.

3.1. Formal definition of Fair-PG

Let $X \in \mathcal{X} \subset \mathbb{R}^d$ be an input feature, $Y \in \mathcal{Y} = \{1, \dots, M\}$ be a target label. We also denote $A \in \mathcal{A} = \{1, \dots, N\}$ as a group label defined by one or multiple sensitive attributes. For example, if phenotype and gender are sensitive attributes, our group labels are $\{$*lighter-skinned male, lighter-skinned female, darker-skinned male* and *darker-skinned female*$\}$. Fair-PG assumes that the input space $\mathcal{X}$ is partitioned into the group-labeled and group-unlabeled sets, $\mathcal{X}_L$ and $\mathcal{X}_U$. That is, a sample $(x, a, y) \sim P(X, A, Y)$ has a group label if $x \in \mathcal{X}_L$, and vice versa if $x \in \mathcal{X}_U$ as illustrated in Fig. 2. With partially annotated group labels, our goal is to find a classifier $f : \mathcal{X} \to \mathcal{Y}$ not biased against the group label A while predicting a target label that best corresponds to an input feature.

3.2. Fairness criterion

Various group fairness criteria have been proposed with different philosophies of how to define discrimination [11, 15, 19]. We consider the *equal opportunity* (EOpp) [19] for M-ary classification problem with non-binary group labels, while most of group fairness criteria assumes binary target or group labels. A classifier f satisfies EOpp with respect to the sensitive group label A and the target Y if the model prediction $\tilde{Y}$ and A are conditionally independent given Y, *i.e.*, $\forall a, a' \in \mathcal{A}$, $y \in \mathcal{Y}$, $P(\tilde{Y} = y | A = a, Y = y) = P(\tilde{Y} = y | A = a', Y = y)$. For measuring the degree of unfairness of f under the distribution $P(X|A, Y)$, we use two types of *disparity of EOpp (DEO)* upon taking the maximum or the average over y as follows, respectively:

$$\begin{aligned} \Delta(f, P, y) :=\Big(\max_{a,a'} \Big(|\mathbb{E}_{P(X|A=a,Y=y)}[\mathbb{I}(f(X) = y)] \\ - \mathbb{E}_{P(X|A=a',Y=y)}[\mathbb{I}(f(X) = y)]| \Big)\Big), \end{aligned} \tag{1}$$

$$\Delta_M(f, P) \triangleq \max_y \Delta(f, P, y), \quad \Delta_A(f, P) \triangleq \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \Delta(f, P, y) \tag{2}$$

The above two metrics indicate the accuracy gap between groups given a target label and complement each other in showing the worst case and average accuracy gaps.

4. Confidence-based Group Label Assignment

In this section, we present our **C**onfidence-based **G**roup **L**abel assignment (**CGL**), which is simple and readily applicable to any fair-training method for the Fair-PG scenario. Our method is described in details, and its theoretical understanding is provided as well.

4.1. Method overview

As the existing fair-training methods for group fairness explicitly utilize group labels to optimize the fairness constraints, they are not directly applicable to our Fair-PG problem. A naive approach to apply the existing methods to Fair-PG is to use only group-labeled samples for the training. Unfortunately, as we observed in Fig. 1 and our experiments, this naive baseline performs even worse than the scratch training method that only uses the target labels, in terms of fairness. As another baseline, we can employ a pseudo-labeling strategy [30] that assigns the estimated group labels to the group-unlabeled data by training a separate group classifier. The vanilla pseudo-labeling strategy enables the recent improvements in algorithmic fairness to be readily transferred into our Fair-PG scenario, other than

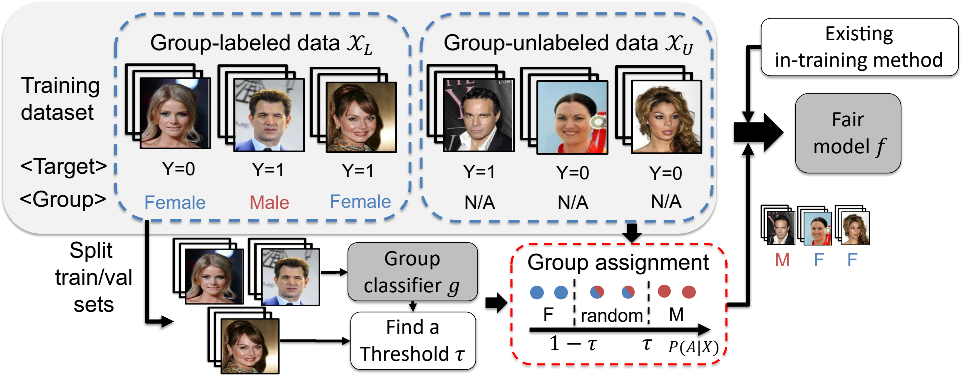


Figure 2. **Overview of the proposed CGL strategy under the Fair-PG scenario.** We train a fair model by assigning group pseudo-labels to group-unlabeled training set $\mathcal{X}_U$. We train an auxiliary group classifier g to generate pseudo labels. Here, we assign random group labels to low confident samples as shown in the dotted red box of the figure ($A = 0$ and $A = 1$ indicates "Female" and "Male", respectively). After assigning confidence-based pseudo-labels to the group-unlabeled training set, we apply a fair-training method to train a fair model f.

the scratch training only with the labeled training set. However, since this vanilla pseudo-labeling strategy trains the group classifier only with the group-labeled training samples, the pseudo group labels can be noisy and incorrect. Compared to the SSL problem, the incorrect group pseudo labels may lead to a more severe issue in terms of fairness by propagating group label errors into the complex fair-training methods.

To that end, we employ **C**onfidence-based **G**roup **L**abel assignment (**CGL**) to reduce the effects of incorrect pseudo-labels. As classifier confidences can be a proxy measure of the mis-classification for the given samples [18, 21], we assume that the low-confident predictions are incorrect. We assign *random* group labels to those less confident group prediction samples, drawn from the empirical conditional distribution of group labels a given the target labels y (*i.e.*, $P(A|Y = y)$) (Line 4 in Algorithm 1). In Section 4.2, we make two theoretical contributions. One is to show that our strategy is better than the vanilla pseudo-labeling with respect to the DEO given metric in Section 2, and the other is to show that the random label assignment is equivalent to ignoring the fairness constraint for those random labeled, low-confident samples. In practice, we expect that our random labeling can play as a regularization method.

For our CGL, we need one hyperparameter, a confidence threshold τ, to determine whether the given prediction is low confident. We split the given group-labeled training set into training and validation sets (Line 1 and 2 in Algorithm 1) and search the best confidence threshold τ satisfying the best accuracy on predicting whether the given prediction is correct or wrong (Line 3 in Algorithm 1). A similar threshold-based strategy is employed in the out-of-distribution sample detection task [21]. As shown in our experiment, there exists a sweet spot of the confidence threshold τ, where $\tau = 1$ is the same as the "random label" assignment to all group-unlabeled samples and $\tau = 0$ is the same as the vanilla pseudo-labeling strategy. Once we have the group classifier and the confidence threshold, we assign the pseudo-group labels with our strategy and train the classifier with base off-the-shelf fair-training method on the pseudo-group labeled training samples. Algorithm 1 and Fig. 2 illustrate the overview of the proposed CGL.

```
Algorithm 1: Confidence-based Group Label assignment (CGL)
Data: Group-labeled training set $\mathcal{X}_L$ and
      group-unlabeled training set $\mathcal{X}_U$.
Result: A set of pseudo group-labels $\tilde{A}$ for
        group-unlabeled training set $\mathcal{X}_U$.
1 Split $\mathcal{X}_L$ into training and validation sets $\mathcal{X}_L^{\text{tr}}$, $\mathcal{X}_L^{\text{val}}$.
2 Train a group classifier $g : \mathcal{X} \to S^{|\mathcal{A}|}$ using the
    training samples $(x, a, y) \sim \mathcal{X}_L^{\text{tr}}$, where $S^{|\mathcal{A}|}$ is
    $|\mathcal{A}|$-simplex.
3 Search a confidence threshold $\tau$ on $\mathcal{X}_L^{\text{val}}$ that satisfies
    $\max_\tau \sum_{x \in \{x | \max g(x) > \tau\}} \mathbb{I}(\arg\max g(x) = a) + \sum_{x \in \{x | \max g(x) \leq \tau\}} \mathbb{I}(\arg\max g(x) \neq a)$.
4 Assign group pseudo-labels $\tilde{a}$ to $(x, y) \sim \mathcal{X}_U$ by
    $\tilde{a} = \arg\max g(x)$ if $\max g(x) > \tau$, otherwise by
    sampling from the empirical conditional
    distribution of $a$ given $y$, i.e., $\tilde{a} \sim P(A|Y = y)$.
```

4.2. Theoretical understanding of CGL

In this subsection, we provide theoretical understandings of why the random label assignment to low confident samples is better than the vanilla pseudo-label (PS) with respect to DEO (Δ). We apply each strategy directly on the true group probability $P(A|X, Y)$, *i.e.*, given an ideally trained group classifier. Our first theoretical result (Proposition 1) supports that the difference between DEO obtained by our CGL and the underlying DEO is smaller than the difference between the DEO obtained by the vanilla strategy and the underlying DEO. In other words, our PS with the random assignment is a better approximation of the true $P(A|X,Y)$ than the vanilla PS. The formal statement is as follows.

Proposition 1. *Assume a binary group $\mathcal{A} = \{0,1\}$. Let $\Delta(x,y;f,P)$ be the influence of x on DEO, $\Delta(f,P)$ (abbreviated as $\Delta(x,y)$). That is, from $\Delta(f,P) = T(\sum_{x\in\{x|f(x)=1\}} |\Delta(x,y)|)$ where $T(\cdot)$ is the maximum or average over y, $\Delta(x,y)$ is defined as follows:*

$$\Delta(x,y) \triangleq P(X=x|A=1,Y=y) - P(X=x|A=0,Y=y).$$

Let $\overline{P}(A|X,Y)$ and $\widehat{P}(A|X,Y)$ be modified distributions by the vanilla pseudo labeling and CGL, respectively:

$$\overline{P}(A=a|X=x,Y=y) = \mathbb{I}\left(P(A=a|X=x,Y=y) > 0.5\right).$$

$$\widehat{P}(A=a|X=x,Y=y)$$
$$= \begin{cases} 1, & \textit{if } P(A=a|X=x,Y=y) \in [\tau,1]. \\ P(A=a|Y=y) & \textit{if } P(A=a|X=x,Y=y) \in (1-\tau,\tau). \\ 0, & \textit{otherwise,} \end{cases}$$

where $0.5 \le \tau \le 1$ is a threshold value. Then, we denote $\overline{P}(X|A,Y)$ and $\widehat{P}(X|A,Y)$ as the distributions induced by $\overline{P}(A|X,Y)$ and $\widehat{P}(A|X,Y)$. We also define $\overline{\Delta}(x,y) \triangleq \Delta(x,y;f,\overline{P})$ and $\widehat{\Delta}(x,y) \triangleq \Delta(x,y;f,\widehat{P})$ as the estimations of $\Delta(x,y)$ Then, for any classifier f, each y and all $x \in \{x|f(x)=1 \text{ and } 1-\tau < P(A=1|X=x,Y=y) < \tau\}$, there exists a τ that satisfies following inequality:

$$|\Delta(x,y) - \overline{\Delta}(x,y)| > |\Delta(x,y) - \widehat{\Delta}(x,y)|. \qquad (3)$$

The proof is in the Appendix. It helps to get a high-level understanding of the advantages of CGL over the vanilla PS although the Proposition 1 is not exactly equivalent to the inequality for $\Delta(f,P)$. In practice, due to the simplicity, we approximate the true distribution $P(A=a|X,Y)$ by one group classifier g which learns $P(A=a|X)$, instead of training $|\mathcal{Y}|$ group classifiers for each y. As reported from our experiment in the Appendix, our group classifier approximates $P(A=a|X)$ well by achieving more than 85% group accuracy with 50% of group-labeled training data.

We also show that assigning random labels to a partition of the given data points, $\mathcal{X}_U$, is equivalent to ignoring the DEO constraint to the data points in $\mathcal{X}_U$:

Proposition 2. *Assume $\mathcal{X}$ is partitioned into any two sets, $\mathcal{X}_L$ and $\mathcal{X}_U$. Let $\widetilde{P}(A|X,Y)$ be a modified version of $P(A|X,Y)$ as follows:*

$$\widetilde{P}(A=a|X=x,Y=y)$$
$$= \begin{cases} P(A=a|X=x,Y=y), & \textit{if } x \in \mathcal{X}_L. \\ P(A=a|Y=y) & \textit{otherwise.} \end{cases}$$

We denote $\widetilde{P}(X|A,Y)$ as a modified data distribution of $P(X|A,Y)$ induced by $\widetilde{P}(A|X,Y)$. Then, for any classifier f, $\Delta(f,P)$ is the same as $\Delta(f,\widetilde{P})$ using the partial set $\mathcal{X}_L$.

The proof is in the Appendix. In practice, since the existing fair-training methods use a relaxed version of DEO, our method can play as a regularization method to the group-unlabeled samples by assigning random group labels.

5. Experiments

In this section, we demonstrate the effectiveness of our CGL for the Fair-PG scenario. We evaluate CGL with various baseline fair-traning methods on three benchmark datasets: UTKFace [44] (the sensitive group is ethnicity), CelebA [31] (the sensitive group is gender) ProPublica COMPAS [23] (the sensitive group is ethnicity), and UCI Adult [14] (the sensitive group is gender) datasets, where COMPAS and Adult datasets are non-vision tabular datasets. We combine CGL with MFD [24], FairHSIC [34] and Re-weighting [22]. To understand the trained group classifier, we provide extensive analysis on the group classifier. Finally, we show the strong empirical contribution of CGL by utilizing extra group-unlabeled training data on the UTKFace dataset. Our CGL shows significant improvements on the target accuracy and the group fairness compared to the baseline methods.

5.1. Experimental settings

5.1.1 Datasets

UTKFace [44]. UTKFace is a facial image dataset, widely adopted as a multi-class and multi-group benchmark. UTKFace contains more than 20K images with annotations, such as age (range from 0 to 116), gender (male and female) and ethnicity (*"White"*, *"Black"*, *"Asian"*, *"Indian"* and *"Others"*). We set ethnicity and age as the sensitive attribute and the target label, respectively. We divided the target age range into three classes: ages between 0 to 19, 20 to 40 and more than 40, following Jung *et al.* [24]. We use four ethnic groups, *"White"*, *"Black"*, *"Asian"* and *"Indian"*, while *"Others"* is excluded. The test set is constructed to contain the same number of samples for each group and target.

CelebA [31]. CelebA contains about 200K face images annotated with 40 binary attributes. As the previous works [24] and [36], we picked up *"Blond Hair"* and *"Attractive"* as the target labels and *"Gender"*(denoted as *"Male"* in the dataset) as the sensitive attribute. The results for "Attractive" and ethical concerns are provided in the Appendix. The test set is constructed as the same as UTKFace.

ProPublica COMPAS [23]. We also consider a non-vision tabular dataset to show the versatility of CGL to other modalities. We use the ProPublica COMPAS dataset, a binary classification task where the target label is whether a defendant reoffends. We set ethnicity as the sensitive attribute and used the same pre-processing as Bellamy *et al*. [4], thereby it includes 5,000 data samples, binary group (*"Caucasian"* and *"Non-Caucasian"* and target labels.

In addition, we also give details and results on **UCI Adult** [14] dataset in the Appendix.

5.1.2 Base fair-training methods

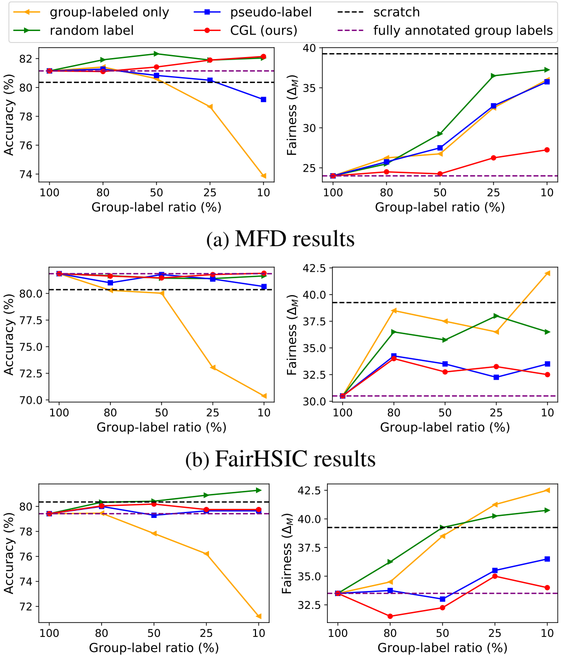
We employ three state-of-the-art in-processing methods, *MMD-based Fair Distillation* (MFD) [24], *FairHSIC* [34] and *Label Bias Correction* (LBC) [22] for the base fair-training method of CGL. We briefly describe each method in the Appendix. We only consider scalable fair-training methods for deep learning-based vision applications; note the primal approaches in group fairness [26, 41] cannot be applied to the vision domain with high dimensional data and complex models (*e.g*., DNNs). We emphasize, however, that our method is not confined to the three methods used in this study but can be easily applied to any fair-training method.

5.1.3 Implementation details

We provide the implementation details in the Appendix, including the details of architectures and optimizers, the hyperparameter search protocol.

Model selection. For fairness-aware learning on real datasets, there can exist a trade-off between accuracy and fairness (see an example of the trade-off in Fig. 6). For fair comparison, we should select the optimal hyperparameters showing the best for one criterion while maintaining similar performance for others. Therefore, we select the hyperparameter showing the best fairness criterion Δ_M while achieving at least 95% of the vanilla training model accuracy for UTKFace and CelebA datasets. We set the lower bound to 90% for the COMPAS dataset. If there exists no hyperparameter achieving the minimum target accuracy, we report the hyperparameter with the best accuracy. All models are chosen from the last training epoch.

Baseline methods and evaluation metrics. The existing in-processing methods for group fairness are not applicable directly to our scenario, *i.e*., when the group labels are not fully annotated. Also, as mentioned in Sec. 2, the existing SSL methods mostly cannot be directly applied to Fair-PG as well, since it is not clear whether they achieve the group fairness when applied to predict non-annotated group labels (see the results of UPS [35] in the Appendix, which is one of the state-of-the-art SSL methods). We hence employ three straightforward baselines for comparison. The **group-labeled only** strategy discards the group-unlabeled samples and only uses the group-labeled samples for the training.



(a) MFD results

(b) FairHSIC results

(c) LBC results

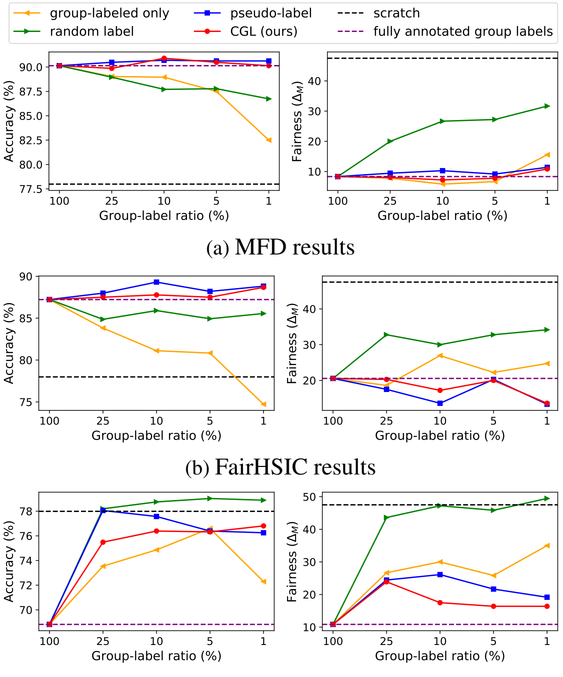
Figure 3. **Results on UTKFace.** For varying group-label ratios in training dataset, we show the combination of three fair-training methods with "group-labeled only" (yellow), "random label" (green), "pseudo-label" (blue) and our CGL (red). "scratch" denotes the vanilla training without a fairness criteria and "fully annotated group labels" denotes the fair-training methods using the full group labels (*i.e*., when group-label ratio is 100%). Higher accuracy and lower Δ_M denote improvements, respectively.

We also examine two group label assignment strategies: The **random label** strategy assigns random labels to all of the group-unlabeled data (drawn from $P(A|Y = y)$), while the **pseudo-label** strategy fully trusts the group classifier predictions. Each method is an extreme case of CGL by setting $\tau = 1$ and $\tau = 0$, respectively. We note that based on the Proposition 2, "random label" has the same effect on evaluating the fairness loss part only with the group-labeled samples while evaluating the main loss with all the samples.

We considered three evaluation metrics for all experiments, the target accuracy, Δ_M and Δ_A (see Eq. (2)). The results are the average scores of four different runs on UTKFace and COMPAS and two different runs on CelebA. Δ_A and standard deviation scores are given in the Appendix.

5.2. Main results

Fig. 3 compares the target accuracies and Δ_M of the combination of MFD, FH and LBC with three baseline strategies and CGL on the UTKFace dataset with different group-label ratios from 100% (fully group annotated) to 10 %. We show the similar results on the CelebA dataset in Fig. 4 where group-label ratio is chosen from 100% to

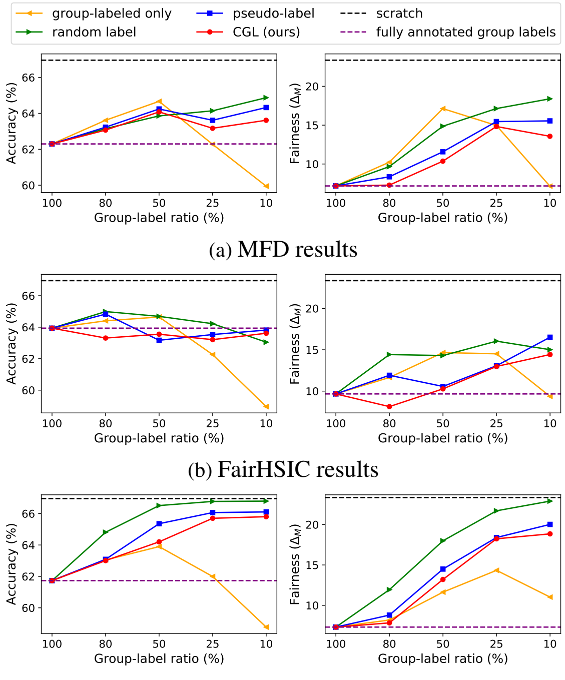


(a) MFD results

(b) FairHSIC results

(c) LBC results

Figure 4. **Results on CelebA.** The details are the same as Fig. 3.



(a) MFD results

(b) FairHSIC results

(c) LBC results

Figure 5. **Results on COMPAS.** The details are the same as Fig. 3.

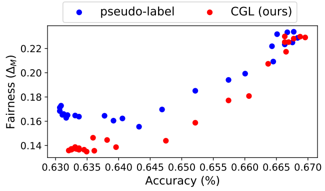


Figure 6. **Accuracy-fairness trade-off on COMPAS with 10% group-labeled training set.** We show accuracy and Δ_M, obtained for CGL and "pseudo label" combined with MFD, for different hyperparameters of MFD..

1%. Note that we choose different group-label ratios to the datasets because UTKFace is multi-class and multi-group dataset, and CelebA is binary-class and binary-group dataset. We also emphasize that the performance comparison of the three baselines and CGL mainly focus on Δ_M because we reported the best Δ_M of each method among models having accuracy more than the accuracy lower bound described in Sec. 5.1.3.

In the figures, the "group-labeled only" (yellow line) consistently shows much worse accuracies than the baseline methods. Especially, when the group-label ratio decreases, "group-labeled only" drastically harms accuracy and fairness at the same time. The "random label" (green line) strategy rarely hurts the accuracies since it uses the full target labels for training, but it shows a drastic drop in Δ_M. The "pseudo-label" (blue line) performs better than the other baselines, but the classifier errors severely affect the fairness performances, especially in the multi-group scenario (*e.g.*, UTKFace). On the other hand, CGL shows consistently better performances than other baselines in most cases, most notably on UTKFace, by successfully handling samples with low confident group predictions.

We also report the results on non-vision tabular dataset in Fig. 5. We observe similar results to Fig. 3 and Fig. 4. Note that "group-labeled only" shows better fairness criterion in price of the rapid decrease of accuracy in the low group label regime. Our method generally performs better than other baselines in all methods in terms of fairness. We point out that although the accuracies of CGL are slightly lower than those of the other baselines, it does not necessarily mean that those schemes perform better since they must sacrifice much more accuracy to achieve similar Δ_M to CGL. To clarify this, we plotted the full accuracy-fairness trade-offs with different hyperparameters on COMPAS in Fig. 6. We clearly observe that CGL dominates "pseudo label" by achieving a better Pareto trade-off curve, which implies the validity of our model selection rule given in Section 5.1.3.

5.3. Analysis of group classifiers

Group classifier confidences. We show the highest and lowest confident samples by the group classifier on UTKFace in Fig. 7. As shown in the figure, the low confident samples are qualitatively uncertain to humans due to diverse lighting, various orientations and low quality. Therefore, our confidence-based thresholding can capture the inherent uncertainty of the dataset. In the Appendix, we provide the

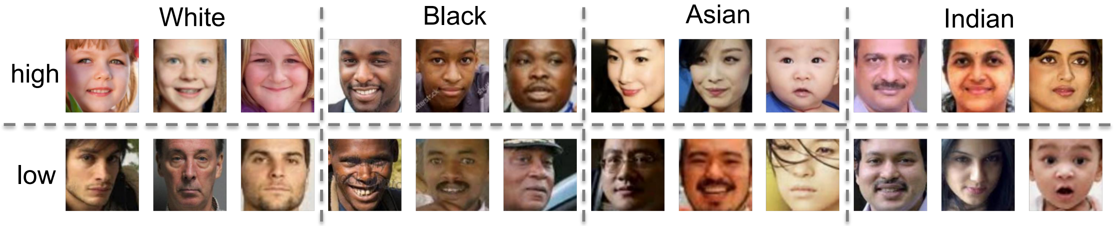


Figure 7. **High and low confident samples by the group classifier on UTKFace.**We illustrate the top-3 highest and lowest confident samples for that the classifier predicts the correct answer from the UTKFace training samples for each group.

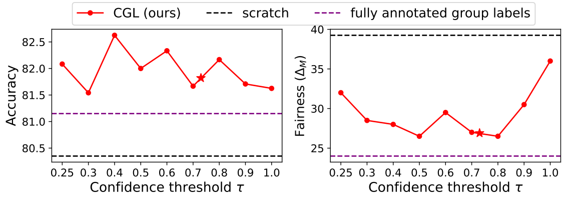


Figure 8. τ **study on UTKFace.** Accuracies and fairness (by Δ_M) for varying τ. $\tau \leq 0.25$ is the same as "random label" and $\tau = 1$ corresponds to vanilla "pseudo-label" in other figures.

confidence score distribution and the group classifier accuracies for different group label ratios.

Study on the threshold τ. Figure 8 shows the accuracies and Δ_M of CGL and MFD by varying τ on UTKFace with 10% group-labeled training set. We fix the hyperparameters used in Fig. 3 and report the average of two different runs. $\tau \leq 0.25$ (since there are four groups) and $\tau = 1$ is equivalent to "random label" and "pseudo-label" in the previous results, respectively. The "$\star$" represents the results of τ achieved by the our strategy (Line 4 in Algorithm 1). Here, we observe that there exists a sweet spot of the threshold that shows better Δ_M and accuracy than "random label" and "pseudo-label", and our method can achieve a good threshold near the sweet spot.

Table 1. **Impact of CGL on UTKFace and extra group-unlabeled training dataset.** The accuracy and fairness criterion on the UTKFace test set are shown. For "MFD + CGL", we assign group psuedo-labels by CGL to the extra group-unlabeled samples from FairFace (73,377 images) with the group classidier trained on full UTKFace training set (20,813 images). We then train MFD on the psuedo-labeled training dataset (94,190 images).

	UTKFace only		UTKFace + FairFace	
	Scratch	MFD	Scratch	MFD + CGL
Accuracy (↑)	80.29	83.46	81.15	**84.38**
Fairness Δ_A (↓)	20.17	16.67	15.67	**13.00**
Fairness Δ_M (↓)	39.00	25.00	24.00	**19.50**

5.4. Augmenting with extra group-unlabeled data

We finally show the impact of the Fair-PG scenario and our CGL on the UTKFace dataset and an extra group-unlabeled dataset. We use the FairFace dataset [27] for the extra dataset. FairFace contains 108,501 facial images with balanced attributes. We filter out ethnicity not in "White", "Black", "Asian" and "Indian". After the filtering, we have 73,377 extra samples. To examine our Fair-PG problem, we let the extra datasets only have the target labels (*i.e.*, ages) but not the group labels. Tab. 1 shows the results of the scratch and MFD trained only on the UTKFace, and the scratch and MFD+CGL on the UTKFace augmented with the FairFace dataset as above. Interestingly, MFD on UTKFace only shows worse fairness ($\Delta_M = 25.0$) than the scratch training on the UTKFace + FairFace ($\Delta_M = 24.0$), which is in line with the result on low group label regime in (Fig. 1). We achieve the state-of-the-art accuracy (84.38%) and fairness ($\Delta_M = 19.5$) by successfully augmenting UTKFace with the extra group-unlabeled dataset.

6. Concluding Remark

We considered a practical learning scenario in which the group labels are partially annotated for fariness-aware learning. We have observed that the existing fair-training method can even perform worse than the scratch training when the number of group labels is small. We propose a simple yet effective solution that is readily applicable to any fair-training method and demonstrated that CGL improves various baselines on several benchmarks. We believe our method can significantly reduce the cost for obtaining additional group labels for all the training samples, facilitating more rapid development of fair classifiers.

Acknowledgments

This work was supported in part by the New Faculty Startup Fund from Seoul National University, NRF Mid-Career Research Program [NRF-2021R1A2C2007884], IITP grants funded by the Korean government [No.2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data], [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], [No.2021-0-02068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], and SNU-NAVER Hyperscale AI Center.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Int. Conf. Mach. Learn.*, pages 60–69. PMLR, 2018. 1

[2] Aditya K. Menon Alex Lamy, Ziyuan Zhong and Nakul Verma. Noise-tolerant fair classification. In *Adv. Neural Inform. Process. Syst.*, volume 32, 2019. 2

[3] Wael Alghamdi, Shahab Asoodeh, Hao Wang, Flavio P Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model projection: Theory and applications to fair machine learning. In *IEEE Int. Sympo. Info. Theory*, pages 2711–2716. IEEE, 2020. 2

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018. 6

[5] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *Int. Conf. Learn. Represent.*, 2019. 3

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Adv. Neural Inform. Process. Syst.*, 2019. 3

[7] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009. 1

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conf. Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018. 1

[9] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Trans. Neural Networks*, 20(3):542–542, 2009. 3

[10] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conf. Fairness, Accountability and Transparency*, pages 339–348, 2019. 2

[11] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. 3

[12] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *Int. Conf. Learn. Represent.*, 2021. 1, 2

[13] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *Int. Conf. Mach. Learn.*, 2019. 2

[14] Dheeru Dua, Casey Graff, et al. Uci machine learning repository. http://archive.ics.uci.edu/ml, 2017. 2, 5, 6

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Conf. Innov. Theor. Compu. Scien.*, 2012. 2, 3

[16] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for fair and efficient machine learning. *Conf. Fairness, Accountability and Transparency*, 2017. 2

[17] Clare Garvie. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016. 1

[18] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Int. Conf. Mach. Learn.*, pages 1321–1330. PMLR, 2017. 4

[19] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Adv. Neural Inform. Process. Syst.*, 2016. 2, 3

[20] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Int. Conf. Mach. Learn.*, pages 1929–1938. PMLR, 2018. 2

[21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Int. Conf. Learn. Represent.*, 2017. 4

[22] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *Int. Conf. Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020. 1, 2, 5, 6

[23] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. There's software used across the country to predict future criminals. and its biased against blacks. *ProPublica*, 2016. 1, 2, 5

[24] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12115–12124, 2021. 1, 2, 5, 6

[25] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 2021. 2

[26] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint Euro. Conf. Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012. 2, 6

[27] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conf. App. Comput. Vis.*, pages 1548–1558, 2021. 2, 8

[28] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010. 1

[29] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Adv. Neural Inform. Process. Syst.*, 33:728–740, 2020. 2

[30] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Int. Conf. Mach. Learn. Worksh.*, 2013. 3

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*, pages 3730–3738, 2015. 2, 5

[32] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 2

[33] Laurent Son Nguyen and Daniel Gatica-Perez. Hirability in the wild: Analysis of online conversational video resumes. *IEEE Trans. Multimedia*, 18(7):1422–1437, 2016. 1

[34] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8227–8236, 2019. 1, 2, 5, 6

[35] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *Int. Conf. Learn. Represent.*, 2021. 3, 6

[36] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *Int. Conf. Learn. Represent.*, 2019. 5

[37] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[38] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Int. Conf. Comput. Vis.*, pages 692–702, 2019. 1

[39] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. In *Adv. Neural Inform. Process. Syst.*, volume 33, pages 5190–5203, 2020. 2

[40] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019. 2

[41] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Int. Conf. Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017. 1, 2, 6

[42] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Int. Conf. Mach. Learn.*, pages 325–333. PMLR, 2013. 1, 2

[43] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conf. AI, Ethics, and Society*, 2018. 2

[44] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5810–5818, 2017. 1, 2, 5