Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2] Taesup Moon[1]
[1] Seoul National University [2] NAVER AI Lab

Abstract

*Recently, fairness-aware learning have become increasingly crucial, but we note that most of those methods operate by assuming the availability of fully annotated 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 **Fair**ness with the **P**artially annotated **G**roup labels (**Fair-PG**). We observe that the existing fairness methods, which only use the data with group-labels, 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. Our 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 for UTKFace, CelebA and COMPAS datasets that by combining CGL and the state-of-the-art fairness-aware in-processing methods, the target accuracies and the fairness metrics are jointly improved compared to the baseline methods. Furthermore, we convincingly show that our CGL enables to naturally augment the given group-labeled dataset with external datasets only with target labels so that both accuracy and fairness metrics can be improved. We will release our implementation publicly to make future research reproduce our results.*

1. Introduction

Recent advances of deep neural networks (DNNs) have witnessed promising outcomes even in societal applications, such as credit estimation [34], crime assessment systems [8, 30], automatic job interviews [40], face recognition [9, 48], and law enforcement [20]. However, DNNs are often more inaccurate to a particular group (*e.g.*, darker-skinned females) than other groups (*e.g.*, lighter-skinned males) [9], *i.e.*, DNNs are discriminatory. This can cause severe societal impacts, *e.g.*, underestimating achievements by students from poorer backgrounds in the UK [3]. To mitigate the issue, *fairness-aware learning* has recently emerged; a model should not discriminate against any group with sensitive attributes, *e.g.*, age, gender, or ethnicity.

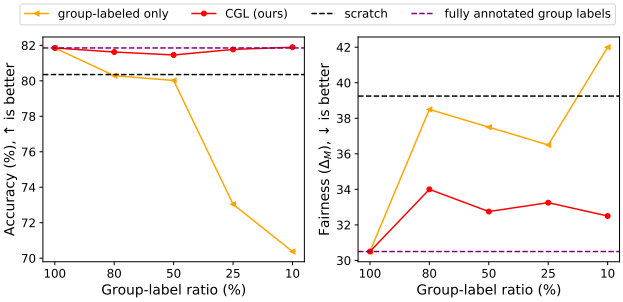


Figure 1. **Can fairness approaches still learn fair classifiers when group labels are partially annotated?** Surprisingly, the state-of-the-art fairness method FairHSIC [42] using only the group-labeled subset (yellow line) shows worse fairness criterion (Δ_M defined in Eq. (1), lower the better) than the "scratch" (*i.e.*, no consideration of a fairness criteria) in the low group label regime (*e.g.*, 10%) on the UTKFace dataset [54]. Our CGL (red line), on the other hand, can be potentially applied to any fairness 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, 13, 29, 31, 42, 50, 51] utilize two types of labels: *target labels* and *group labels*. Target labels are task-oriented (*e.g.*, crime assessment) and group labels are defined by sensitive attributes (*e.g.*, ethnicity). The existing methods rely on 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 computer vision applications, assuming 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 [8, 30, 34] but are defined in high-level semantics, requiring additional expensive human annotations.

*Works done while doing an internship at NAVER AI Lab.

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. From these reasons, the underlying assumption by the previous fairness methods, *i.e.*, group labels are fully annotated, can limit their usability in real-world applications.

Contribution. In this work, to the best of our knowledge, we propose and investigate a less explored but very practical problem for the first time: *Algorithmic **Fair**ness with the **P**artially annotated **G**roup labels (**Fair-PG**)*. Since the existing fairness methods need group labels for all training samples to optimize fairness constraints, they cannot be directly applied to the Fair-PG problem. We empirically show that the existing fairness methods perform even worse than the vanilla training baseline without fairness criteria in terms of fairness when the number of group labels is small (*e.g.*, when only 10% training samples have group labels) – See Fig. 1. As the second contribution, we propose a simple yet effective baseline strategy to enable *any* fairness methods can be applied to Fair-PG, named **C**onfidence-based **G**roup **L**abel assignment (**CGL**). CGL assigns pseudo group labels to group-unlabeled samples by an auxiliary group classifier if the predictions are sufficiently confident. Otherwise, for low confident samples, we assign random group labels.

We provide high-level understandings of how CGL works on the Fair-PG scenario. We theoretically support that 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 group classifier. We also theoretically show that assigning a random group label to a data point implies the elimination of the fairness constraint of the sample. In practice, since the existing fairness 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 fairness methods (*e.g.*, [31], FairHSIC [42] and Re-weighting [29]) has consistently and significantly improved the target accuracy and the fairness parity even under the low group label regime on facial image datasets [39, 54] and tabular datasets [30]. For example, compared to the "group-labeled only" baseline, the combination of MFD and CGL shows +8.23% target accuracy and -8.75 disparity of equalized odds (DEO) on UTKFace [54] when only 10% data points have group labels. Furthermore, we remarkably outperform MFD trained only on UTKFace by +0.92% accuracy and -5.5 DEO measures on UTKFace dataset by utilizing extra group-unlabeled facial images [33] by CGL.

2. Related Works

Algorithmic fairness. Several fairness criteria have been proposed in various aspects, including disparate impact [18], equalized odds [26], fairness through unawareness [17] and counterfactual fairness [37]. Under the fairness criteria, many fair-training methods have been developed to mitigate unfairness of a model. The fairness methods can be divided into three categories depending on where the technique for fairness is injected into; *pre-processing* methods [16,42,51] fix a training dataset before learning a model; *in-processing* [13, 29, 31, 32, 50, 52] methods consider fairness during training time; *post-processing* methods [2] modify a trained model. We emphasize again that despite technical advances for achieving group fairness, existing methods for group fairness have not considered the situation that a partition of a training dataset has no group label (*i.e.*, Fair-PG). Through the rest of the paper, we show that CGL can be easily applied into existing fairness methods and prevent their performance degradation under Fair-PG.

Biases in machine learning. Emerging studies on DNNs have revealed that DNNs rely on shortcut biases [4, 10, 21, 22, 44]. The existing de-biasing methods let a model less attend on the dataset biases in an implicit way by using extra biased networks [4, 10] or data augmentations [22] without using bias labels. Both fairness methods and de-biasing methods aim to learn a representation invariant to undesired decision cues, such as sensitive groups and dataset biases. However, de-biasing methods explore implicit shortcut biases that harm the network generalizability, where many known shortcuts (*e.g.*, language bias [10] or texture bias [22]) are neither strongly relative to ethical concerns nor easy to configure. On the other hand, in the fairness problem, sensitive groups are diversely defined by the target application to avoid negative societal impacts (*i.e.*, a model should make the same predictions to any social group such as ethnicity or gender). Therefore, even though de-biasing methods can be applied to Fair-PG by ignoring group labels, there is no guarantee to learn fair models by the de-biasing approaches. In this work, we focus on fairness methods explicitly utilizing group labels for the base method of CGL.

Semi-supervised learning. Semi-supervised learning (SSL) [11] 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, while SSL only aims to the target task, we aim to learn a fair classifier. In addition, the recent state-of-the-art SSL methods [6, 7, 46] 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 [38] to avoid complex modifications on the base fairness methods by assigning pseudo group labels to group-unlabeled samples. While the original PL fully trusts the network predictions, we set the random labels to the low confident samples. We theoretically show that our random label selection strategy shows better fairness constraints than the vanilla PL. Our strategy is also similar to the recently proposed UPS [43] which withdraws pseudo-labels for low confident and high uncertain samples by using an external predictive uncertainty module. However, CGL uses the full group-unlabeled samples by the random label strategy.

3. Problem Definition

In this section, we formally define our target scenario, Fair-PG, and the fairness criterion, *disparity of equalized odds (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 [12, 17, 26]. In this work, we consider the *equalized odds* (EO) [26] because most of group fairness criteria including EO assume that target label or group label is binary, while EO can be easily extended to the M-ary classification problem with non-binary group labels. A classifier f satisfies EO 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 EO (DEO)*, $\Delta(f, P)$, upon taking the maximum or the average over y as follows, respectively:

$$\Delta_M(f, P) \triangleq \max_y \Big(\max_{a,a'} \Big(|\mathbb{E}_{P(X|A=a,Y=y)}[\mathbb{I}(f(X) = y)] - \mathbb{E}_{P(X|A=a',Y=y)}[\mathbb{I}(f(X) = y)]| \Big) \Big), \tag{1}$$

$$\Delta_A(f, P) \triangleq \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \Big(\max_{a,a'} \Big(|\mathbb{E}_{P(X|A=a,Y=y)}[\mathbb{I}(f(X) = y)] - \mathbb{E}_{P(X|A=a',Y=y)}[\mathbb{I}(f(X) = y)]| \Big) \Big). \tag{2}$$

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 method, named **C**onfidence-based **G**roup **L**abel assignment (**CGL**), that is simple and readily applicable to any fairness method to our Fair-PG scenario. Our method assigns pseudo-labels to the high confident samples and random labels to the low confident samples. We also provide theoretical understanding of how CGL works; our assignment strategy is better than the vanilla pseudo-labeling strategy with respect to the fairness constraint. Also we show that the random labeling to a sample implies ignoring fairness constraint of the sample.

4.1. Method overview

As the existing fairness methods 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 [38] 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 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 [25, 28], we assume that the low-confident predictions are incorrect. We

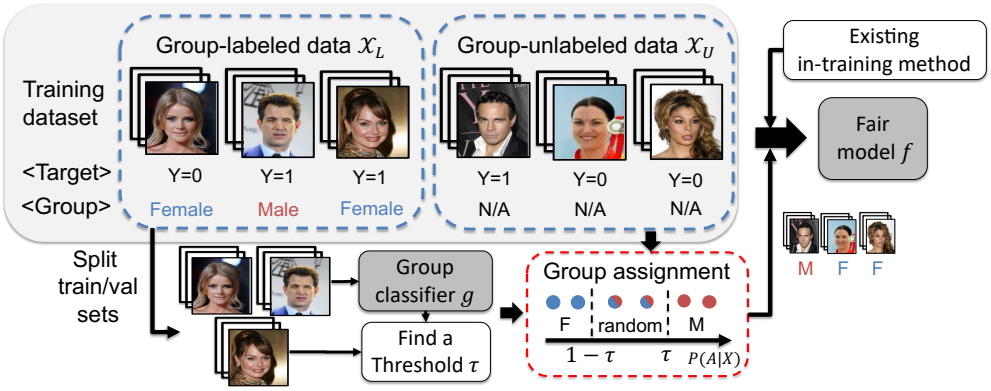


Figure 2. **Overview of the proposed CGL strategy under the Fair-PG scenario.** Under the partially annotated group labels, we train a fair model by assigning group pseudo-labels to group-unlabeled training set $\mathcal{X}_U$. We train an auxiliary group classifier g to generate pseudo labels. Here,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

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 strategy with respect to the *disparity of equalized odds*, 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 [28]. 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 fairness method on the pseudo-group labeled training samples. Algorithm 1 and Fig. 2 illustrate the overview of the proposed CGL.

Algorithm 1: Confidence-based **G**roup **L**abel 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 τ on $\mathcal{X}_L^{\text{val}}$ that satisfies $\max_\tau \sum_{x \in \{x | \max g(x) > \tau\}} \mathbb{I}(\arg\max g(x) = a) + \sum_{x \in \{x | \max g(x) \le \tau\}} \mathbb{I}(\arg\max g(x) \neq a)$.
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 with respect to the *disparity of equalized odds* (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 pseudo-label strategy with the random assignment is a better approximation of the true $P(A|X, Y)$ than the vanilla pseudo-labeling strategy. 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:

$$\overline{P}(A = a|X = x, Y = y) = \begin{cases} 1, & \text{if } P(A = a|X = x, Y = y) > 0.5. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

$$\widehat{P}(A = a|X = x, Y = y) = \begin{cases} 1, & \text{if } P(A = a|X = x, Y = y) \in [\tau, 1]. \\ P(A = a|Y = y) & \text{if } P(A = a|X = x, Y = y) \in (1 - \tau, \tau). \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $0.5 \leq \tau \leq 1$ 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 $x \in \{x|f(x) = 1$ and $1 - \tau < P(A = 1|X = x, Y = y) < \tau\}$, there exists a τ that satisfies following inequality:

$$|\Delta(x, y) - \overline{\Delta}(x, y)| > |\Delta(x, y) - \widehat{\Delta}(x, y)|. \tag{5}$$

The proof of the proposition is given in Appendix A. It helps to get a high-level understanding of the advantages of our CGL over the vanilla pseudo labeling 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 $|\mathcal{Y}|$ number of group classifiers for each y. As reported from our experiment in Appendix C, our group classifier approximates $P(A = a|X)$ well by achieving more than 85% with 50% of group-labeled training data.

We also present a conceptual intuition of the random label assignment of CGL. We show that assigning random labels to a partition of the given data points, $\mathcal{X}_U$, is equivalent to ignoring the DEO fairness constraint to the data points in $\mathcal{X}_U$ by the following proposition.

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:*

$$\widetilde{P}(A = a|X = x, Y = y) = \begin{cases} P(A = a|X = x, Y = y), & \text{if } x \in \mathcal{X}_L. \\ P(A = a|Y = y) & \text{otherwise.} \end{cases} \tag{6}$$

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 again given in Appendix A. In practice, since the existing fairness methods use the relaxed version of DEO (*e.g.*, by the finite-sample estimator of Hilbert-Schmidt Independence Criterion (HSIC)), our method can play as a regularization method to the group-unlabeled samples by assigning them to 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 fairness methods on three benchmark datasets: UTKFace [54] (the sensitive group is ethnicity), CelebA [39] (the sensitive group is gender) and ProPublica COMPAS [30] datasets (the sensitive group is ethnicity), where the COMPAS dataset is a non-vision tabular dataset. We combine CGL with MFD [31], FairHSIC [42] and Reweighting [29]. 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 [54]. 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*. [31]. 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 [39]. CelebA contains about 200K face images annotated with 40 binary attributes. As Jung *et al*. [31], we picked up *"Attractive"* as the target label and *"Gender"* (denoted as *"Male"* in the dataset) as the sensitive attributes by considering the ratio of balance for each attribute. The test set is constructed as the same as UTKFace.

ProPublica COMPAS [30]. 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*. [5], thereby it includes 5,000 data samples, binary group (*"Caucasian"* and *"Non-Caucasian"* and target labels.

5.1.2 Base fairness methods

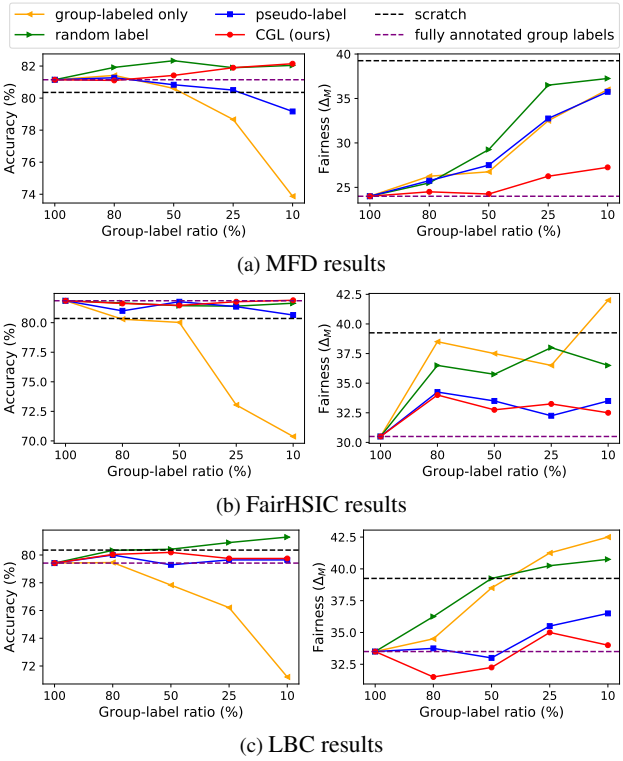
We employ four state-of-the-art in-processing methods, *MMD-based Fair Distillation* (MFD) [31], *FairHSIC* [42] and *Label Bias Correction* (LBC) [29] for the base fairness method of CGL. We briefly describe each method in the Appendix B.3. We only consider scalable fairness methods to deep learning based vision applications. For example, the primal approaches in group fairness [32, 50] cannot be applied into the vision domain with high dimensional data or complex models (*e.g.*, DNNs). However, we emphasize that our method is not limited to the four methods used in this study but can be easily applied to any fairness method.

5.1.3 Implementation details

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

Model selection. Fairness-aware learning has two contradictory goals, accuracy and fairness. For a fair comparison, we should select the optimal hyperparmeters 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. 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 comparison methods and evaluation metrics. The existing fair-training methods are not available to be applied to our scenario, *i.e.*, when the group labels are not fully annotated. To the best of our knowledge, our study is the first work to solve the Fair-PG problem, and hence, we employ three straightforward baselines for comparison. The **group-labeled only** strategy withdraws the group-unlabeled samples and only uses the group-labeled samples for the training. We also examine two group label assignment strategies: The **random label** strategy assigns random labels to 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 considered three evaluation metrics for all experiments, the target accuracy, Δ_M Eq. (1) and Δ_A Eq. (2), and due to page limitation, we report Δ_A in the Appendix E. The results are the average scores of four different runs on UTKFace and COMPAS and two different runs on CelebA. The standard deviation scores are also in Appendix E.



(a) MFD results

(b) FairHSIC results

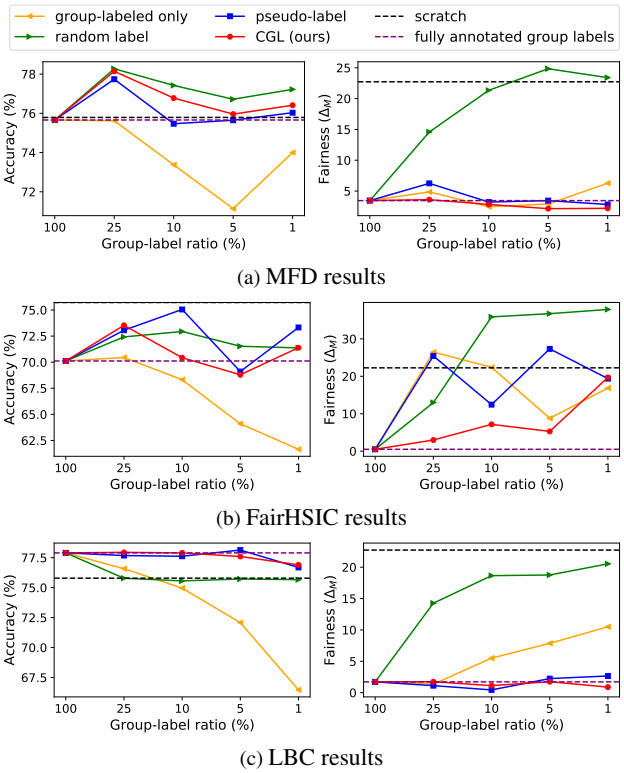
(c) LBC results

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

5.2. Main results

Fig. 3 compares the target accuracies and Δ_M of the combination of MFD, FH and LBC with three baselines 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% (fully group annotated) to 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 the 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

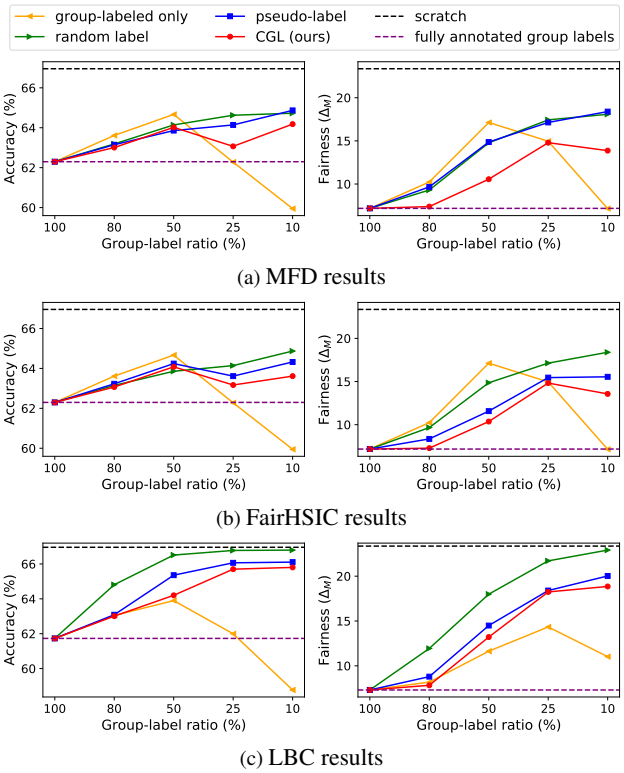


(a) MFD results

(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.

line) strategy rarely hurts the accuracies since it uses the full target labels for training, but it shows a drastic drop in Δ_M. "pseudo-label" (blue line) performs better than the other baselines. However, the classifier errors severely affect the fairness performances, especially in a multi-group scenario (*e.g.*, on UTKFace). On the other hand, our CGL shows consistently better performances than other baselines in most cases. Particularly on UTKFace, CGL consistently outperforms all methods 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 because of the rapid decreases of accuracies 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 less than the other baselines, it does not mean a performance gain of baselines since they must sacrifice much more accuracy to achieve similar Δ_M to CGL by our model selection rules. However, we note that our method can be limited in this scenario because learning a group classifier from the tabular data could occur a severe overfitting issue. We expect that the overall results can be improved by enhancing the group classifier performances.

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. 6. 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 Appendix C, we provide the confidence score distribution and the group classifier accuracies for different group label ratios.

Study on the threshold τ. Figure 7 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. $\tau \leq 0.25$ (because the number of groups is four) and $\tau = 1$ are equivalent to "random label" and "pseudo-label" in the previous results, respectively. Here, we observe again that there exists a sweet spot of the threshold that shows better Δ_M and accuracy than "random label" and vanilla "pseudo-label".

5.4. Towards SOTA fairness-aware classifiers 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-

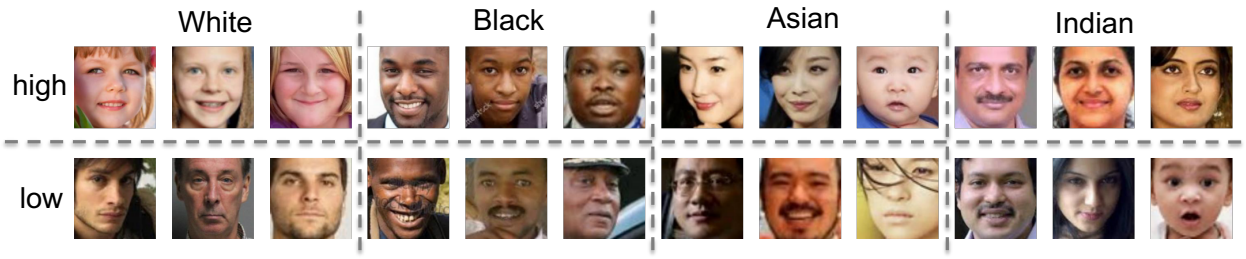


Figure 6. **High and low confident samples by the group classifier on UTKFace.**We illustrate the top-3 highest and lowest confident samples for that the classifier predicts the correct answer from the UTKFace training samples for each group.

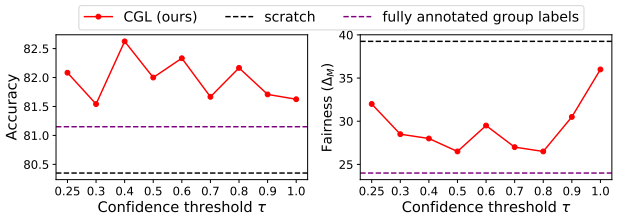


Figure 7. τ **study on UTKFace.** Accuracies and fairness (by Δ_M) by varying threshold τ are shown. Here, $\tau \leq 0.25$ is the same as "random label" and $\tau = 1$ is the same as vanilla "pseudo-label" in other figures. Note that $\tau \leq 0.25$ will show the same results to $\tau = 0$ because there are four groups in the dataset.

unlabeled dataset. We use the FairFace dataset [33] 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 filtration, we have 73,377 extra samples. To examine our Fair-PG problem, we let the extra datasets only have target labels (*i.e.*, ages) but not group labels. Tab. 1 shows the results of the scratch and MFD trained only on the UTKFace dataset, and the scratch and MFD+CGL on the UTKFace augmented with group-unlabelled FairFace dataset. Interestingly, MFD on UTKFace only shows worse fairness criterion (25.0 Δ_M) than the scratch training on the UTKFace + FairFace (24.0 Δ_M), which is in line with the low group label regime in (Fig. 1). We achieve the state-of-the-art accuracy (84.38%) and fairness criterion (19.5) by successfully augmenting UTKFace with the extra group-unlabeled dataset.

6. Concluding Remark

We have delved into an overlooked but practical learning scenario when 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, *e.g.*, the scratch training on group-labeled UTKFace combined with group-unlabeled FairFace outperforms MFD trained only on UTKFace in fairness. We propose a simple yet effective solution that is readily applicable to any fairness method. We have demonstrated that CGL improves various base fairness methods on several fairness benchmarks. While DNNs can cause negative societal impacts by dismissing fairness criterion, we expect our CGL can mitigate the negative societal impacts economically by only annotating a subset of group-unlabeled samples. We strongly encourage to investigate our Fair-PG scenario by fairness researchers to make fairness method practical. We conclude the paper with final remark on the limitation.

Table 1. **Impact of CGL on UTKFace and extra group-unlabeled training dataset.** The accuracies 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) using the group labels of the full UTKFace training set (20,813 images) and 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**

Limitations. Although our method can be applied to any fairness method, we observe that CGL is not always better than other baselines. First, our method relies on the quality of group classifier, hence, if the group classifier performs worse, our method does not guarantee better fairness than the vanilla pseudo-labeling. Also, the group classifier predictions can be noisy. In Appendix, we show group prediction accuracy of our group classifier. In the low group label regime, the accuracy of our classifier decreases to less than 80% on UTKFace. This implies that if the base method is sensitive to noisy group labels (*e.g.*, Adversairal De-biasing), our method and pseudo-labeling can perform worse than our expectation.

Acknowledgement

We thank to NAVER AI Lab colleagues to provide valuable comments and early stage reviews, especially to Song Park, Sangdoo Yun and Dongyoon Han. We also thank to Seoul National University colleagues to detailed discus-

sions and reviews, particularly to Donggyu Lee and Taeeon Park.

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] 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 *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2711–2716. IEEE, 2020. 2

[3] An algorithm determined uk students' grades. chaos ensued. https://www.wired.com/story/an-algorithm-determined-uk-students-grades-chaos-ensued/. 1

[4] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *Int. Conf. Mach. Learn.*, 2020. 2

[5] 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. 5

[6] 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 *ICLR*, 2019. 2

[7] 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. 2

[8] 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

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 1

[10] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *Adv. Neural Inform. Process. Syst.*, pages 839–850, 2019. 2

[11] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 2

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

[13] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *ICLR*, 2021. 1, 2

[14] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 14

[15] Sanghyuk Chun, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2019. 14

[16] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning (ICML)*, 2019. 2

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Conference on Innovations in Theoretical Computer Science (ITCS)*, 2012. 2, 3

[18] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015. 2

[19] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 14

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

[21] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[22] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2018. 2

[23] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 12

[24] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005. 12

[25] 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. 3, 14

[26] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 3

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 12

[28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 3, 4

[29] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020. 1, 2, 5, 6, 13

[30] 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, 11

[31] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *CVPR*, pages 12115–12124, 2021. 1, 2, 5, 6, 12, 13

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

[33] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 2, 8, 11

[34] 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

[35] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 11

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 12

[37] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[38] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 3

[39] 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*, pages 3730–3738, 2015. 2, 5, 11

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

[41] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *International Conference on Learning Representations (ICLR)*, 2019. 14

[42] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8227–8236, 2019. 1, 2, 5, 6, 12, 13

[43] 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. *arXiv preprint arXiv:2101.06329*, 2021. 3, 14

[44] Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will dnns choose? a study from the parameter-space perspective. *arXiv preprint arXiv:2110.03095*, 2021. 2

[45] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 14

[46] 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. 2

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 14

[48] 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 *ICCV*, pages 692–702, 2019. 1

[49] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 14

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

[51] 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

[52] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018. 2, 12

[53] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2017. 14

[54] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017. 1, 2, 5, 11

Supplementary Materials

We include additional materials in this document. We first state our societal impact and dataset license in the beginning. We provide a detailed proof of our propositions in Appendix A. We include our implementation details, such as architecture, optimization, hyperparameter search and base fairness methods and their modifications in Appendix B. We provide the additional results in Appendix E.

Societal impact. As we stated in the main text, a vanilla DNN training can occur negative societal impacts by dismissing fairness criterion, on the other hand, considering fairness criterion at the training time requires a huge number of group labels. We expect our CGL can bridge the gap between real-world applications and fairness-aware training, so that mitigating the negative societal impacts economically by only annotating a subset of group-unlabeled samples.

Dataset license. In the paper, we use four datasets: UTKFace [54], CelebA [39], ProPublica COMPAS [30] and FairFace [33]. According to the official web page[1], UTKFace dataset is a non-commercial license dataset where the copyright belongs to the original owners in the web. The dataset is built by Dlib [35] and annotations are tagged by the DEX algorithm and human annotators. CelebA dataset has a similar license statement[2] to UTKFace. COMPAS dataset is collected its data points from Broward County Sheriff's Office in Florida[3] which is a public records. FairFace is licensed by CC by 4.0[4]. Overall, all datasets have clean licenses that is applicable to any public research project.

A. Proof of propositions

A.1. Proof of Proposition 1

Proof. We only show only the case where $P(A=1|X=x,Y=y) \geq 0.5$ and the opposite case can be proved in the same way. For any classifier f and all $x \in \{x|f(x)=1 \text{ and } 0.5 \leq P(A=1|X=x,Y=y) < \tau\}$, we have from $\overline{P}$ and $\widehat{P}$ defined in (Eq. (3) and (4), manuscript),

$$\overline{\Delta}(x,y) = \big(\frac{1}{P(A=1|Y=y)}\big)P(X=x|Y=y), \tag{A.1}$$

$$\widehat{\Delta}(x,y) = 0. \tag{A.2}$$

Then, we have

$$|\Delta(x,y) - \overline{\Delta}(x,y)| - |\Delta(x,y) - \widehat{\Delta}(x,y)| = \begin{cases} \overline{\Delta}(x,y) & \text{if } \Delta(x,y) \leq 0 \\ \overline{\Delta}(x,y) - 2\Delta(x,y) & \text{otherwise.} \end{cases} \tag{A.3}$$

For the first case in Eq. (A.3), we can trivially see that $\Delta(x,y) > 0$. For the second case in Eq. (A.3), we have

$$\overline{\Delta}(x,y) - 2\Delta(x,y) \tag{A.4}$$

$$= \big(\frac{1-2P(A=1|X=x,Y=y)}{P(A=1|Y=y)} + \frac{2P(A=0|X=x,Y=y)}{P(A=0|Y=y)}\big)P(X=x|Y=y) > 0 \tag{A.5}$$

, if $P(A=1|X=x,Y=y) < \frac{P(A=1|Y=y)+1}{2}$. Therefore, we have the proposition 1 by setting τ to $\frac{P(A=1|Y=y)+1}{2}$. □

A.2. Proof of Proposition 2

Proof. Given a data distribution $P(X,A,Y)$ and a classifier f, $\Delta(f,P)$ is defined as follows:

$$\Delta(f,P) = T\Big(\max_{a,a'}\big(|\underbrace{P(\hat{Y}=y|A=a,Y=y) - P(\hat{Y}=y|A=a',Y=y)}_{(a)}|\big)\Big) \tag{A.6}$$

[1] https://susanqq.github.io/UTKFace/
[2] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[3] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[4] https://github.com/joojs/fairface

, where $T(\cdot)$ can be the maximum or average over y depending on the types of Δ. For each y, a and a', the above argument of $\max_{a,a'}$, (a) in Eq. (A.6), can be represented as follows:

$$\begin{aligned}(a) &= \sum_{x\in\{x|f(x)=y\}} P(X=x|A=a,Y=y) - P(X=x|A=a',Y=y) \\ &= \sum_{x\in\{x\in X_L|f(x)=y\}} P(X=x|A=a,Y=y) - P(X=x|A=a',Y=y) \\ &\quad + \sum_{x\in\{x\in X_U|f(x)=1\}} P(X=x|A=a,Y=y) - P(X=x|A=a',Y=y) \end{aligned} \tag{A.7}$$

Then, the second term of Eq. (A.7) can be represented as follows:

$$\begin{aligned}&\sum_{x\in\{x\in X_U|f(x)=y\}} P(X=x|A=a,Y=y) - P(X=x|A=a',Y=y) \\ &= \sum_{x\in\{x\in X_U|f(x)=y\}} \frac{P(A=a|X=x,Y=y)P(X=x|Y=y)}{P(A=a|Y=y)} - \frac{P(A=a'|X=x,Y=y)P(X=x|Y=y)}{P(A=a'|Y=y)}\end{aligned} \tag{A.8}$$

If we substitute $P(A|X,Y)$ into $\widehat{P}(A|X,Y)$ in the RHS of Eq. (A.8), we have the proposition 2. □

B. More Implementation Details

B.1. Architecture and optimization

We choose the same architecture for the base classifier and the group classifier; ResNet18 [27] for the UTKFace and CelebA experiments and a simple 2-layered neural network for the COMPAS experiments. On UKTFace and CelebA datasets, we train the models with the Adam optimizer [36] for 70 epochs by setting the initial learning rate 0.001 reduced by 0.1 when the loss is stagnated for 10 epochs following Jung *et al*. [31]. We train the model for 50 epochs on COMPAS dataset. All results are reported by the model at the last epoch.

B.2. Hyperparameter search

In the experiments, there are two types of hyperparameters: the confidence threshold of CGL, and the method-specific hyperparameters for each method. Since our method only needs the group-labeled training dataset for training group classifier and seeking a threshold, we split the group-labeled samples into 80% training and 20% validation samples. The confidence threshold is searched on the validation set (by Algorithm 1, manuscript).

Fairness-aware training methods are usually sensitive to the hyperparameter selection due to the accuracy-fairness trade-off; when the strength for fairness is getting stronger, the target accuracy is getting worse. For example, a trivial solution to achieve the fairest classifier is to predict all labels to a constant label, while this solution is the worst solution in terms of the target accuracy. Hence, the careful tuning of the control parameters to fairness criteria (*e.g*., MMD [31], HSIC [42] or adversarial loss [52]) takes the key role in handling the accuracy-fairness trade-off. In our experiments, we aim to find a fair classifier while showing *a comparable accuracy* to the vanilla training method. Thus, we select the hyperparameter showing the best fairness criterion Δ_M while achieving at least 95% of the vanilla training model accuracy. We set the lower bound to 90% for the COPMAS dataset. If there exists no hyperparameter achieving the minimum target accuracy, we report the hyperparameter with the best accuracy. We perform the grid search on the hyperparameter candidates for every partial group-label case and for every method. The full hyperparameter search space is illustrated in Tab. B.1.

B.3. Base fairness methods and their modifications

Here, we describe the overview of each base fairness method used for the experiments. MFD and FairHSIC use additional fairness-aware regularization terms as the relaxed version of the targeted fairness criteria. MFD proposed a *maximum-mean-discrepancy*-based [23] regularization term to achieve fairness via feature distillation and FairHSIC devised a *HSIC*-based [24] regularization term to obtain feature representations independent on group labels. For FairHSIC, we only implement the second term of their decomposition loss (*i.e*., the HSIC loss between the feature representations and the group labels).

LBC is a re-weighting algorithm optimizing weights of examples through multiple iterations of full training to ensure their theoretical guarantees. The original LBD requires multiple full-training iterations by alternatively computing a EO criterion

Method	Hyperparameter	Candidates
MFD [31]	MMD strength λ	[10, 30, 100, 300, 1000, 3000, 10000, 30000]
FairHSIC [42]	HSIC strength λ	[1,3,10, 30, 100, 300, 1000, 3000]
LBC [29]	Adversary strength α learning rate of adversary	[1, 3, 10, 30, 100] $[10^{-4}, 10^{-2}]$

Table B.1. **Hyperparameter search spaces.** We perform the grid search on the validation set to find the best hyperparameters for each method. We use the same hyperparameters for optimizer (See Appendix B.1).

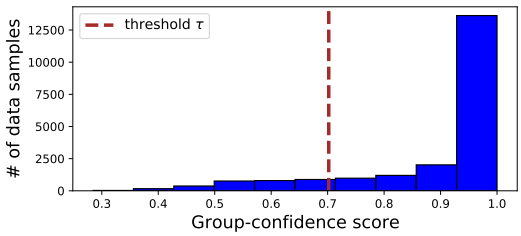


Figure C.1. **Group confidences verse sample densities.** The number of samples for each confidence bin is shown. The red dotted line denotes the selected threshold in the UTKFace experiments.

Table C.1. **Group classifier performances.** We compare the accuracies by the baseline decision rule ($\arg\max$) and by our method (assigning random labels to low confident samples) for the trained group classifiers on the small group-labeled training samples.

Group-label ratio	80%	50%	25%	10%
Baseline	87.88	86.11	82.82	77.73
Ours	87.24	85.81	82.59	75.21

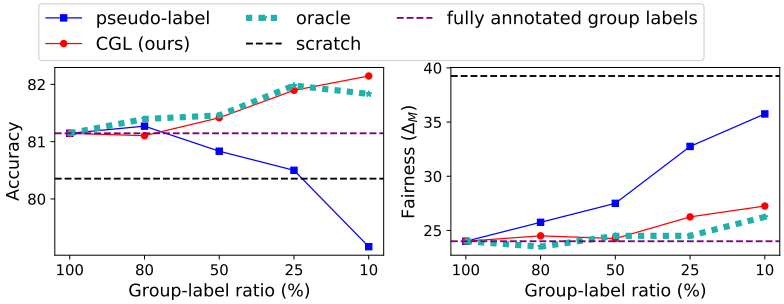


Figure C.2. **Comparisons with an "oracle" fair group classifier on UTKFace with MFD.** The oracle classifier group classifier has the same accuracy with our group classifier (used for "pseudo-label" and "CGL (ours)" – See Tab. C.1) but the wrong samples by the "oracle" classifier are *randomly* chosen from the dataset.

after full-training and re-training the full dataset by optimal weights. This alternative optimization needs a very huge training budget. We modify the EO computation iteration to a few-epoch iterations, *i.e.*, 5 epochs, instead of the full-training.

AD lets an adversary cannot predict group labels by the additional adversarial loss. In our experiments, AD shows little improvements if the group or target label is not binary where Jung *et al.* [31] witnessed the same phenomenon. Thus, we use multiple adversaries for AD to make AD be available to solve multi-class and multi-group problems following Jung *et al.* [31] and omit the loss projection in the original objectives of AD for a stable learning. Also, we only report AD results for the Compas dataset while AD does not perform well on other vision datasets.

Table D.1. **Accuracy on COPMAS for AD.**

	100%	80%	50%	25%	10%
group-labeled only	63.51 (±1.45)	65.32 (±0.58)	63.65 (±0.37)	61.30 (±1.22)	57.52 (±2.84)
random label		63.61 (±0.55)	63.11 (±0.67)	64.44 (±1.38)	64.67 (±0.24)
psuedo-label		64.55 (±0.41)	64.12 (±0.63)	63.19 (±0.18)	65.80 (±0.38)
CGL		63.05 (±1.13)	63.25 (±0.60)	64.24 (±1.24)	63.82 (±1.55)

Table D.2. Δ_A **on COPMAS for AD.**

	100%	80%	50%	25%	10%
group-labeled only	10.35 (±1.84)	13.32 (±2.14)	11.46 (±0.63)	9.75 (±1.84)	5.27 (±0.76)
random label		9.26 (±1.46)	10.69 (±1.46)	13.17 (±2.10)	11.90 (±1.44)
psuedo-label		12.43 (±3.39)	12.11 (±4.07)	11.37 (±3.17)	16.26 (±0.57)
CGL		9.63 (±3.60)	11.93 (±3.90)	14.71 (±1.27)	10.67 (±2.70)

C. Additional Analysis of Group Classifiers

Prediction confidences by our group classifier. In the main manuscript, we show the highest and lowest confident samples by the group classifier on UTKFace in Fig. 6. As shown in the figure, low confident samples are qualitatively uncertain to humans due to diverse lighting, various orientations and low quality, where Shi *et al*. observed the same results by an uncertainty-aware face embedding [45]. From the qualitative results, we observe that our confidence-based threshold method can reasonably capture the inherent uncertainty of the dataset without an explicit uncertainty-aware training, such as MC-Dropout [19] or probabilistic embeddings [14, 41].

However, because our group classifier does not guarantee to capture proper uncertainty measures, we presume that applying an uncertainty-aware training can improve CGL as Rizve *et al*. [43]. We show the number of samples by the confidences in Fig. C.1. Our classifier shows high confident predictions (over 65% predictions are confident than 0.9 because) because it is not trained by calibration-aware regularizations [25] or other regularization techniques known to help confidence calibration scores [15], such as mixed sample augmentations [49, 53] and smoothed labels [47]. Nonetheless, we observe that many images are still low confident and our group classifier can figure whether the prediction is correct or wrong; when we apply the optimal threshold, our classifier has 85.43% accuracy to figure out whether the prediction is wrong or correct.

Quality of our group classifier and the threshold-based decision rule. In Tab. C.1, we show the group accuracies of our group classifier by different decision rules on varying group label ratio. We show two different decision rules: the baseline $\arg\max$ strategy and our confidence-based random altering (*i.e*., $\arg\max$ if the confidence is larger than τ, otherwise $P(A|Y)$) with the best threshold. We observe that our random label strategy slightly hurts the accuracies but not significantly. In other words, our group classifier has well-sorted confidences that can capture the self predictive uncertainty.

Finally, we compare our group classifier and the "oracle" group classifier which has the same accuracy to ours, but group labels that our group classifier wrongly predict are replaced into a group label sampled from an uniform distribution. In other words, "oracle" assumes the scenario where our confidence-based thresholding perfectly operates. Fig. C.2 shows the comparison of CGL, "pseudo-label" and "oracle" on UTKFace dataset and MFD. Here, we see that "oracle" significantly improve the performance in terms of fairness other than "pseudo-label". This imply that only random-labeling for wrongly predicted group labels can prevent performance degradation of DEO, which experimentally supports our proposition 2. We also observe that the performance of CGL is comparable one of "oracle", meaning that random labeling low confident samples are more critical to the performance than high confident samples with noisy group labels.

D. AD results on COMPAS dataset

Tab. D.1, Tab. D.2 and Tab. D.3 compare the target accuraies, Δ_A and Δ_M of the combinations of AD with three baselines and CGL on COMPAS dataset. The number in the parentheses with $\pm$ stands for the standard deviation of each metric obtained several independent runs with different seeds. our CGL again show the better performances than other baselines in terms of fairness for most cases. Through the case where the group-label ratio is 25%, we can see that confidence-based thresholding by a group classifier can be slightly sensitive in the group label regime if the base fairness method is vulnerable

Table D.3. Δ_M **on COPMAS for AD.**

	100%	80%	50%	25%	10%
group-labeled only	12.72 (±2.98)	16.30 (±2.41)	14.39 (±1.50)	12.61 (±2.11)	8.52 (±2.22)
random label		12.37 (±2.09)	13.51 (±1.38)	15.96 (±1.93)	15.70 (±2.26)
psuedo-label		16.15 (±3.79)	15.68 (±4.73)	13.97 (±2.67)	19.57 (±0.93)
CGL		13.78 (±5.00)	14.73 (±5.28)	17.96 (±0.31)	13.23 (±3.82)

Table E.1. **Accuracy on UTKFace for MFD.**

	100%	80%	50%	25%	10%
group-labeled only	81.15 (±0.28)	81.42 (±0.39)	80.60 (±0.37)	78.67 (±0.64)	73.88 (±0.78)
random label		81.92 (±0.36)	82.33 (±0.53)	81.90 (±0.63)	82.04 (±0.34)
psuedo-label		81.27 (±0.60)	80.83 (±0.39)	80.50 (±0.54)	79.17 (±0.54)
CGL		81.10 (±0.24)	81.42 (±0.42)	81.90 (±0.41)	82.15 (±0.58)

Table E.2. Δ_A **on UTKFace for MFD.**

	100%	80%	50%	25%	10%
group-labeled only	15.67 (±0.71)	16.33 (±0.85)	17.08 (±1.46)	18.50 (±1.38)	21.25 (±2.66)
random label		16.83 (±0.29)	18.58 (±0.83)	22.58 (±0.86)	23.50 (±1.80)
psuedo-label		16.33 (±0.97)	16.67 (±0.41)	18.58 (±1.95)	20.00 (±2.16)
CGL		15.33 (±1.03)	14.92 (±2.17)	17.17 (±1.57)	17.25 (±1.04)

Table E.3. Δ_M **on UTKFace for MFD.**

	100%	80%	50%	25%	10%
group-labeled only	24.00 (±1.58)	26.25 (±3.56)	26.75 (±2.59)	32.50 (±2.87)	36.00 (±2.92)
random label		25.50 (±1.66)	29.25 (±4.66)	36.50 (±0.50)	37.25 (±3.19)
psuedo-label		25.75 (±2.86)	27.50 (±0.87)	32.75 (±3.83)	35.75 (±4.49)
CGL		24.50 (±2.06)	24.25 (±2.17)	26.25 (±3.49)	27.25 (±2.77)

Table E.4. **Accuracy on UTKFace for FairHSIC.**

	100%	80%	50%	25%	10%
group-labeled only	81.85 (±0.23)	80.29 (±0.64)	80.02 (±1.10)	73.04 (±3.68)	70.38 (±1.27)
random label		81.67 (±0.48)	81.44 (±0.78)	81.40 (±0.78)	81.65 (±0.56)
psuedo-label		81.00 (±1.02)	81.77 (±0.26)	81.35 (±0.56)	80.65 (±0.59)
CGL		81.62 (±0.79)	81.46 (±0.72)	81.77 (±0.57)	81.90 (±0.89)

to noisy group labels (*e.g.*, AD).

E. More Experimental Results

Table from E.1 to E.27 show the detailed results including accuracy, Δ_A and Δ_M for all experiments in Figure 3, 4 and 5 in the main manuscript. The details of numbers in parentheses are the same as tables in Appendix D.

Table E.5. Δ_A **on UTKFace for FairHSIC.**

	100%	80%	50%	25%	10%
group-labeled only	18.50 (±1.67)	21.33 (±1.62)	21.67 (±1.67)	22.08 (±2.18)	27.42 (±4.30)
random label		22.50 (±1.71)	22.50 (±1.30)	23.75 (±2.17)	23.50 (±1.34)
psuedo-label		21.92 (±1.01)	21.08 (±2.25)	19.75 (±1.77)	20.67 (±0.94)
CGL		20.67 (±1.70)	20.75 (±1.09)	20.42 (±1.11)	18.50 (±1.46)

Table E.6. Δ_M **on UTKFace for FairHSIC.**

	100%	80%	50%	25%	10%
group-labeled only	30.50 (±4.33)	38.50 (±2.96)	37.50 (±3.84)	36.50 (±2.18)	42.00 (±3.67)
random label		36.50 (±3.04)	35.75 (±3.27)	38.00 (±3.67)	36.50 (±2.60)
psuedo-label		34.25 (±3.27)	33.50 (±1.50)	32.25 (±4.97)	33.50 (±1.66)
CGL		34.00 (±3.08)	32.75 (±2.28)	33.25 (±2.86)	32.50 (±2.69)

Table E.7. **Accuracy on UTKFace for LBC.**

	100%	80%	50%	25%	10%
group-labeled only	79.42 (±0.74)	79.46 (±1.16)	77.83 (±0.28)	76.21 (±0.63)	71.21 (±1.06)
random label		80.33 (±0.69)	80.42 (±0.64)	80.90 (±0.62)	81.29 (±0.82)
psuedo-label		80.00 (±0.50)	79.29 (±0.96)	79.65 (±0.97)	79.65 (±0.96)
CGL		80.04 (±0.82)	80.19 (±0.35)	79.75 (±0.74)	79.75 (±0.67)

Table E.8. Δ_A **on UTKFace for LBC.**

	100%	80%	50%	25%	10%
group-labeled only	18.75 (±1.04)	19.58 (±2.95)	21.58 (±1.66)	22.58 (±1.04)	24.67 (±2.25)
random label		19.42 (±0.76)	21.00 (±0.97)	23.08 (±0.86)	22.17 (±1.07)
psuedo-label		19.08 (±1.16)	19.17 (±1.17)	19.75 (±1.93)	19.92 (±1.99)
CGL		18.00 (±2.90)	17.92 (±1.66)	17.83 (±1.83)	19.25 (±1.64)

Table E.9. Δ_M **on UTKFace for LBC.**

	100%	80%	50%	25%	10%
group-labeled only	33.50 (±2.69)	34.50 (±3.84)	38.50 (±1.12)	41.25 (±3.96)	42.50 (±7.09)
random label		36.25 (±1.09)	39.25 (±2.77)	40.25 (±1.92)	40.75 (±2.95)
psuedo-label		33.75 (±2.17)	33.00 (±2.00)	35.50 (±3.35)	36.50 (±2.87)
CGL		31.50 (±5.12)	32.25 (±1.64)	35.00 (±3.32)	34.00 (±1.87)

Table E.10. **Accuracy on CelebA for MFD.**

	100%	25%	10%	5%	1%
group-labeled only	75.66 (±0.12)	75.62 (±0.33)	73.37 (±0.10)	71.14 (±1.23)	74.00 (±0.59)
random label		78.27 (±0.10)	77.42 (±0.44)	76.72 (±0.22)	77.22 (±0.22)
psuedo-label		77.74 (±0.80)	75.47 (±0.13)	75.65 (±0.31)	76.03 (±0.43)
CGL		78.14 (±0.72)	76.78 (±0.12)	75.96 (±0.44)	76.41 (±0.14)

Table E.11. Δ_A **on CelebA for MFD.**

	100%	25%	10%	5%	1%
group-labeled only	2.44 (±0.69)	3.25 (±1.55)	1.61 (±0.40)	2.13 (±0.72)	3.54 (±0.93)
random label		11.22 (±0.25)	15.95 (±0.07)	18.12 (±0.40)	17.97 (±1.07)
psuedo-label		5.68 (±2.34)	2.46 (±0.16)	2.36 (±0.51)	1.99 (±0.10)
CGL		3.30 (±0.09)	2.25 (±1.49)	2.00 (±0.12)	1.96 (±0.31)

Table E.12. Δ_M **on CelebA for MFD.**

	100%	25%	10%	5%	1%
group-labeled only	3.45 (±1.04)	4.86 (±1.78)	2.46 (±1.10)	2.90 (±1.44)	6.27 (±1.15)
random label		4.58 (±0.10)	21.34 (±0.29)	24.84 (±0.44)	23.41 (±1.20)
psuedo-label		6.24 (±2.80)	3.21 (±0.03)	3.45 (±0.31)	2.82 (±0.21)
CGL		3.61 (±0.31)	2.82 (±2.04)	2.14 (±0.26)	2.19 (±0.21)

Table E.13. **Accuracy on CelebA for FairHSIC.**

	100%	25%	10%	5%	1%
group-labeled only	70.11 (±2.69)	67.05 (±2.44)	68.32 (±0.97)	64.09 (±0.89)	61.65 (±0.43)
random label		72.42 (±0.16)	72.94 (±1.23)	71.53 (±0.33)	71.37 (±0.07)
psuedo-label		73.06 (±0.87)	67.29 (±5.04)	69.10 (±0.84)	73.33 (±1.07)
CGL		73.53 (±1.74)	67.51 (±0.93)	68.80 (±3.06)	67.03 (±5.07)

Table E.14. Δ_A **on CelebA for FairHSIC.**

	100%	25%	10%	5%	1%
group-labeled only	0.31 (±0.10)	6.48 (±2.09)	19.11 (±7.72)	6.82 (±1.83)	15.20 (±2.12)
random label		8.82 (±6.91)	24.41 (±0.67)	24.92 (±0.08)	25.50 (±3.37)
psuedo-label		16.78 (±5.81)	2.32 (±1.28)	17.80 (±1.92)	12.30 (±8.96)
CGL		1.76 (±0.61)	2.39 (±1.40)	3.08 (±0.13)	1.92 (±0.27)

Table E.15. Δ_M **on CelebA for FairHSIC.**

	100%	25%	10%	5%	1%
group-labeled only	0.50 (±0.18)	10.76 (±1.99)	22.41 (±6.06)	8.78 (±1.04)	16.85 (±3.53)
random label		12.98 (±10.58)	35.87 (±1.18)	36.73 (±0.37)	37.85 (±3.74)
psuedo-label		25.44 (±8.73)	2.51 (±1.15)	27.32 (±3.34)	19.38 (±12.96)
CGL		2.98 (±1.10)	4.36 (±2.48)	5.28 (±0.10)	2.74 (±0.39)

Table E.16. **Accuracy on CelebA for LBC.**

	100%	25%	10%	5%	1%
group-labeled only	77.90 (±0.12)	76.58 (±0.43)	74.96 (±1.21)	72.08 (±1.20)	66.46 (±2.32)
random label		75.78 (±0.37)	75.56 (±0.48)	75.72 (±0.42)	75.67 (±0.48)
psuedo-label		77.68 (±0.38)	77.62 (±0.24)	78.13 (±0.01)	76.68 (±0.52)
CGL		77.95 (±0.01)	77.90 (±0.14)	77.60 (±0.14)	76.90 (±0.91)

Table E.17. Δ_A **on CelebA for LBC.**

	100%	25%	10%	5%	1%
group-labeled only	1.02 (±0.47)	1.10 (±0.63)	4.36 (±0.13)	5.55 (±2.76)	8.58 (±5.63)
random label		11.42 (±1.70)	14.86 (±2.01)	15.54 (±1.41)	16.54 (±1.96)
psuedo-label		1.04 (±0.26)	0.30 (±0.01)	1.49 (±0.21)	1.65 (±0.37)
CGL		1.40 (±0.25)	0.68 (±0.18)	1.36 (±0.37)	0.59 (±0.01)

Table E.18. Δ_M **on CelebA for LBC.**

	100%	25%	10%	5%	1%
group-labeled only	1.72 (±0.84)	1.36 (±0.78)	5.51 (±0.55)	7.86 (±2.69)	10.50 (±5.12)
random label		14.26 (±2.40)	18.65 (±2.35)	18.76 (±2.14)	20.53 (±2.30)
psuedo-label		1.12 (±0.18)	0.44 (±0.03)	2.25 (±0.10)	2.66 (±0.89)
CGL		1.75 (±0.50)	1.12 (±0.29)	1.75 (±0.24)	0.89 (±0.16)

Table E.19. **Accuracy on COPMAS for MFD.**

	100%	80%	50%	25%	10%
group-labeled only	62.30 (±0.37)	63.61 (±0.45)	64.67 (±0.49)	62.28 (±1.33)	59.95 (±1.55)
random label		63.15 (±0.74)	63.86 (±0.90)	64.14 (±0.70)	64.87 (±0.66)
psuedo-label		63.23 (±0.48)	64.24 (±0.78)	63.61 (±1.27)	64.32 (±0.51)
CGL		63.07 (±0.68)	64.08 (±0.59)	63.17 (±0.68)	63.61 (±1.22)

Table E.20. Δ_A **on COPMAS for MFD.**

	100%	80%	50%	25%	10%
group-labeled only	6.52 (±0.97)	8.57 (±0.34)	13.59 (±2.08)	11.72 (±0.90)	5.13 (±1.13)
random label		7.57 (±1.48)	11.72 (±0.66)	12.84 (±1.67)	14.15 (±1.21)
psuedo-label		6.88 (±0.92)	8.95 (±1.02)	11.09 (±1.80)	12.87 (±1.68)
CGL		6.27 (±1.08)	7.99 (±0.65)	10.70 (±1.90)	10.82 (±2.18)

Table E.21. Δ_M **on COPMAS for MFD.**

	100%	80%	50%	25%	10%
group-labeled only	7.18 (±0.89)	10.24 (±1.14)	17.13 (±2.64)	14.96 (±2.40)	7.15 (±0.69)
random label		9.67 (±3.05)	14.86 (±0.56)	17.13 (±2.68)	18.39 (±2.58)
psuedo-label		8.35 (±1.97)	11.57 (±0.88)	15.46 (±2.12)	15.55 (±2.73)
CGL		7.28 (±1.66)	10.36 (±0.54)	14.82 (±2.60)	13.57 (±4.15)

Table E.22. **Accuracy on COPMAS for FairHSIC.**

	100%	80%	50%	25%	10%
group-labeled only	63.94 (±0.36)	64.40 (±0.70)	64.65 (±0.31)	62.26 (±1.12)	58.95 (±1.46)
random label		64.99 (±0.24)	64.69 (±1.18)	64.22 (±0.66)	63.05 (±0.94)
psuedo-label		64.83 (±0.28)	63.17 (±0.26)	63.53 (±0.64)	63.82 (±0.65)
CGL		63.31 (±0.64)	63.55 (±0.51)	63.21 (±0.33)	63.61 (±0.82)

Table E.23. Δ_A **on COPMAS for FairHSIC.**

	100%	80%	50%	25%	10%
group-labeled only	7.63 (±1.20)	9.80 (±1.21)	11.65 (±2.14)	11.32 (±1.16)	6.59 (±1.90)
random label		11.66 (±1.25)	11.05 (±1.88)	11.91 (±1.90)	11.74 (±1.49)
psuedo-label		9.92 (±1.24)	7.76 (±1.26)	9.91 (±1.85)	11.57 (±1.21)
CGL		6.01 (±1.71)	8.12 (±1.32)	9.37 (±2.11)	10.63 (±1.85)

Table E.24. Δ_M **on COPMAS for FairHSIC.**

	100%	80%	50%	25%	10%
group-labeled only	9.66 (±1.46)	11.65 (±2.01)	14.66 (±2.37)	14.51 (±1.73)	9.36 (±2.67)
random label		14.42 (±2.78)	14.30 (±1.39)	16.04 (±1.78)	15.01 (±3.32)
psuedo-label		11.91 (±2.04)	10.56 (±1.01)	13.07 (±1.38)	16.51 (±2.68)
CGL		8.13 (±3.01)	10.27 (±1.89)	12.98 (±2.84)	14.43 (±3.13)

Table E.25. **Accuracy on COPMAS for LBC.**

	100%	80%	50%	25%	10%
group-labeled only	61.73 (±0.12)	63.05 (±0.21)	63.90 (±0.95)	61.99 (±1.47)	58.77 (±1.31)
random label		64.81 (±0.25)	66.51 (±0.44)	66.77 (±0.30)	66.79 (±0.14)
psuedo-label		63.09 (±0.90)	65.36 (±0.27)	66.07 (±0.39)	66.11 (±0.93)
CGL		63.01 (±0.83)	64.20 (±1.41)	65.70 (±0.28)	65.80 (±1.08)

Table E.26. Δ_A **on COPMAS for LBC.**

	100%	80%	50%	25%	10%
group-labeled only	4.36 (±0.69)	6.05 (±1.37)	8.94 (±1.72)	11.31 (±0.42)	7.61 (±0.60)
random label		9.01 (±0.99)	14.39 (±1.28)	17.70 (±0.74)	18.93 (±0.56)
psuedo-label		5.59 (±1.28)	11.20 (±0.91)	14.70 (±1.53)	16.80 (±1.04)
CGL		4.99 (±1.48)	10.32 (±1.91)	14.24 (±0.74)	15.56 (±1.63)

Table E.27. Δ_M **on COPMAS for LBC.**

	100%	80%	50%	25%	10%
group-labeled only	7.30 (±1.04)	8.18 (±1.57)	11.63 (±1.92)	14.33 (±1.44)	11.02 (±2.31)
random label		11.94 (±1.32)	17.99 (±1.79)	21.71 (±0.98)	22.91 (±1.21)
psuedo-label		8.79 (±1.59)	14.49 (±1.44)	18.40 (±1.68)	20.02 (±2.46)
CGL		7.83 (±2.35)	13.21 (±2.91)	18.24 (±0.42)	18.85 (±2.50)