Learning Fair Classifiers with Partially Annotated Group Labels — Supplementary Materials —

Sangwon Jung Sanghyuk Chun Taesup Moon

Supplementary Materials

We include additional materials in this document. We first state our societal impact, dataset license, limitations and ethical concerns in the beginning. We provide additional related works for biases in machine learning in Appendix B and a detailed proof of our propositions in Appendix A. We include our implementation details, such as architecture, optimization, hyper-parameter search and base fairness methods and their modifications in Appendix C. We provide the additional analysis of group classifiers in Appendix D, experimental results in Appendix E and result tables in Appendix E.5.

Dataset license. In the paper, we use four datasets: UTKFace [36], CelebA [23], ProPublica COMPAS [18] and FairFace [20]. 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 [21] 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.

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.

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. Finally, in the case that a distribution shift for the sensitive attribute exists when predicting group labels of group-unlabeled data from group-unlabeled data, the naive application of would suffer from performance degradation. These distribution shift can be alleviated by training a group classifier with robust optimization techniques (*e.g.*, choosing a distribution shift-aware optimizer [5], invariant risk minimization [1] or group distributed robust optimization [28]).

Ethical concerns We originally used a subjective and potentially unethical "Attractive" attribute in our experiments with the CelebA dataset. It is known that "Attractive" is highly correlated to gender ("Male"), while most other attributes are not [32]. Our purpose of CelebA experiments is to show the scalability of our method as CelebA (200K) is a large-scale dataset compared to UTK (20K), COMPAS (5K), Adult (40K). From a similar motivation, many previous studies employed Attractive as their target label [6, 19, 25, 26]. Particularly, Quadrianto et al. used Attractive "as the proxy measure of getting invited for a job interview in the world of fame" [26]. However, we agree that using a subjective attribute as "Attractive" can

[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

be unethical. We only used the results as an example, and we alert that such classifiers for attractiveness can cause potential ethical concerns.

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:

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

, 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. □

Method	Hyperparameter	Candidates
MFD [19]	MMD strength λ	[10, 30, 100, 300, 1000, 3000, 10000, 30000]
FairHSIC [26]	HSIC strength λ	[1,3,10, 30, 100, 300, 1000, 3000]
LBC [17]	Adversary strength α learning rate of adversary	[1, 3, 10, 30, 100] $[10^{-4}, 10^{-2}]$

Table C.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 C.1).

B. Additional Related Works for Biases in Machine Learning

Emerging studies on DNNs have revealed that DNNs rely on shortcut biases [2, 4, 11, 12, 29]. The existing de-biasing methods let a model less attend on the dataset biases in an implicit way by using extra biased networks [2, 4] or data augmentations [12] 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 [4] or texture bias [12]) 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.

C. More Implementation Details

C.1. Architecture and optimization

We choose the same architecture for the base classifier and the group classifier; ResNet18 [16] 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 [22] 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*. [19]. We train the model for 50 epochs on COMPAS dataset. All results are reported by the model at the last epoch.

C.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 [19], HSIC [26] or adversarial loss [34]) 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. C.1.

C.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 [13] regularization term to achieve fairness via feature distillation and FairHSIC devised a *HSIC*-based

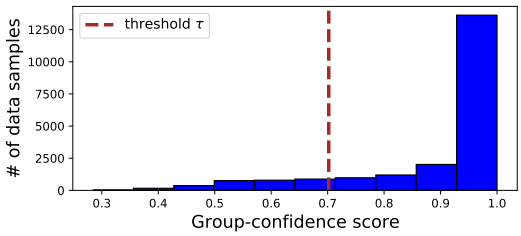


Figure D.1. **Group confidences verse sample densities.** The number of samples for each confidence bin is shown. The red dotted line denotes the selected threshold in the UTKFace experiments.

Table D.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

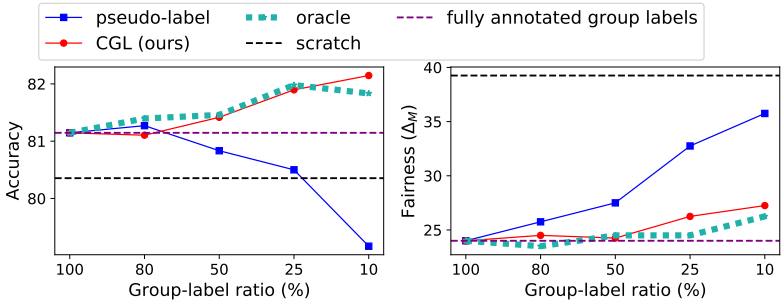


Figure D.2. **Comparisons with an "oracle" fair group classifier on UTKFace with MFD.** The oracle classifier group classifier has the same accuracy with our group classifier (used for "pseudo-label" and "CGL (ours)" – See Tab. D.1) but the wrong samples by the "oracle" classifier are *randomly* chosen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

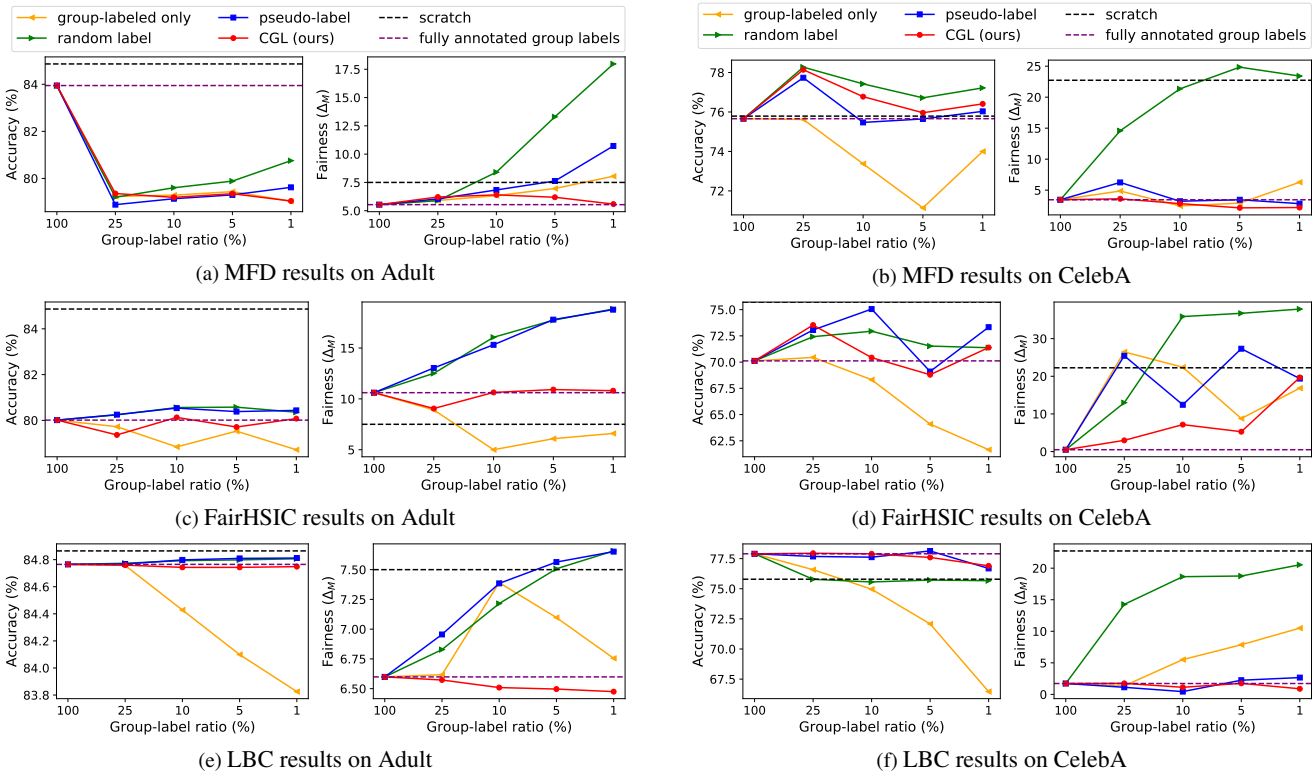
[14] regularization term to obtain feature representations independent on group labels. For FairHSIC, we only implement the second term of their decomposition loss (*i.e.*, the HSIC loss between the feature representations and the group labels).

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 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.* [19] 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.* [19] 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.

D. 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. 7. 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 [30]. 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 [10] or probabilistic embeddings [7, 24].



(a) MFD results on Adult

(b) MFD results on CelebA

(c) FairHSIC results on Adult

(d) FairHSIC results on CelebA

(e) LBC results on Adult

(f) LBC results on CelebA

Figure E.1. **Results on Adult and CelebA. The target label in CelebA is "Attractive" attribute.** The details are the same as Fig. 3.

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*. [27]. We show the number of samples by the confidences in Fig. D.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 [15] or other regularization techniques known to help confidence calibration scores [8], such as mixed sample augmentations [33, 35] and smoothed labels [31]. 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. D.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. D.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.

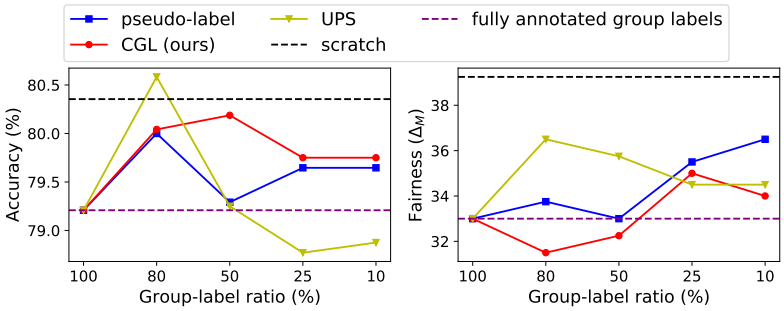


Figure E.2. LBC results on UTKFace

E. Additional experimental results

E.1. Results on Adult dataset

To show the consistent improvements on another dataset, we conducted an additional experiment on Adult dataset with the same details as the main experiment in the manuscript. UCI Adult dataset [9] is a non-vision tabular dataset used for a binary classification task where the target label is whether the income exceeds $50K per a year given attributes about the person. We set gender as the sensitive attribute and used the same processing as Bellamy *et al*. [3], so that it includes 45,000 data samples.

The left column of Fig. E.1 shows the results of CGL and baselines combined with base fairness methods on Adult dataset, and we observe the consistent trend of CGL that our method mostly performs better than other baselines for fairness. We repeatedly note that our slightly lower accuracies do not imply the ineffectiveness of CGL because we report the model with the best DEO where accuracy is lower-bounded.

E.2. Results on CelebA using the "Attractive" attribute as the target label

The right column of Fig. E.1 shows the target accuracy and Δ_M on CelebA using "Attractive" attribute as the target label. From the right column of Fig. E.1, we again demonstrate the better performance of CGL than other baselines for all base fairness methods. Since the "Attractive" attribute would be the subjective and potentially unethical to discuss the results rigorously, as described in the beginning of Appendix, we advise that these results should used only as a auxiliary and not as a primary result.

E.3. Comparison CGL with UPS

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. To corroborate our finding, we carried out additional experiments with a state-of-the-art SSL method, UPS, utilized for Fair-PG. UPS iteratively trains the group classifier and predicts the missing group labels in the training set and filters out the samples with uncertain predictions. (We omitted the negative learning of UPS since it cannot be applied any base fairness methods.) Note such filtering would unnecessarily discard significant amount of the target label information, hence, the accuracy would hurt particularly when the group label ratio is low. In Fig. E.2, we report the result of LBC on UTKFace, including the UPS baseline. We indeed observe that UPS suffers from low accuracy especially when the group-label ratio is low, and CGL mostly outperforms UPS for both accuracy and fairness. This confirms that a naive plug-in of SSL method for Fair-PG would not be satisfactory.

E.4. AD results on COMPAS dataset

Tab. E.1, Tab. E.2 and Tab. E.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 to noisy group labels (*e.g.*, AD).

Table E.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 E.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)

Table E.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.4. **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.5. 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)

E.5. Result tables

Table from E.4 to E.30 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 E.4.

Table E.6. 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.7. **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)

Table E.8. 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.9. 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.10. **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.11. 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.12. 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.13. **Accuracy on CelebA for MFD.**

	100%	25%	10%	5%	1%
group-labeled only	90.14 (±0.12)	89.03 (±0.28)	88.96 (±0.49)	87.50 (±0.42)	82.50 (±2.08)
random label		88.96 (±0.07)	87.71 (±0.21)	87.78 (±0.69)	86.74 (±0.07)
psuedo-label		90.49 (±0.49)	90.69 (±0.28)	90.62 (±0.07)	90.62 (±0.07)
CGL		89.86 (±0.14)	90.90 (±0.07)	90.49 (±0.07)	90.14 (±0.28)

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

	100%	25%	10%	5%	1%
group-labeled only	5.28 (±0.69)	4.72 (±0.56)	4.03 (±0.69)	3.61 (±0.00)	8.61 (±0.00)
random label		11.53 (±0.14)	15.97 (±0.69)	16.39 (±0.56)	18.47 (±0.97)
psuedo-label		5.42 (±0.69)	5.28 (±0.56)	5.14 (±0.42)	6.25 (±0.14)
CGL		5.28 (±0.83)	4.03 (±0.14)	4.58 (±0.42)	6.39 (±0.56)

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

	100%	25%	10%	5%	1%
group-labeled only	8.33 (±1.04)	7.78 (±1.67)	5.83 (±0.83)	6.67 (±0.00)	15.56 (±0.56)
random label		20.00 (±0.00)	26.67 (±2.22)	27.22 (±1.11)	31.67 (±1.11)
psuedo-label		9.44 (±0.00)	10.28 (±1.39)	9.17 (±1.39)	11.39 (±0.28)
CGL		8.06 (±0.83)	7.22 (±0.56)	7.78 (±0.00)	10.83 (±1.39)

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

	100%	25%	10%	5%	1%
group-labeled only	87.22 (±0.42)	83.82 (±0.07)	81.11 (±1.39)	80.83 (±1.94)	74.72 (±1.25)
random label		84.86 (±0.14)	85.90 (±0.21)	84.93 (±0.35)	85.56 (±0.14)
psuedo-label		87.99 (±0.76)	89.31 (±0.69)	88.19 (±0.42)	88.82 (±0.49)
CGL		87.50 (±1.11)	87.78 (±1.39)	87.50 (±1.11)	88.68 (±0.07)

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

	100%	25%	10%	5%	1%
group-labeled only	12.50 (±1.11)	10.42 (±3.75)	15.83 (±2.50)	12.50 (±3.61)	14.72 (±2.78)
random label		20.00 (±1.11)	18.19 (±0.42)	19.31 (±0.14)	20.00 (±0.83)
psuedo-label		10.14 (±3.19)	9.17 (±0.28)	12.50 (±0.28)	7.92 (±0.97)
CGL		11.67 (±1.11)	10.28 (±4.17)	12.22 (±1.11)	9.31 (±2.36)

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

	100%	25%	10%	5%	1%
group-labeled only	20.56 (±1.67)	18.61 (±6.39)	26.94 (±3.61)	22.22 (±6.11)	24.72 (±1.94)
random label		32.78 (±2.78)	30.00 (±1.67)	32.78 (±0.00)	34.17 (±1.39)
psuedo-label		17.50 (±4.72)	13.61 (±0.83)	20.28 (±1.39)	13.33 (±1.67)
CGL		20.28 (±4.17)	17.22 (±7.78)	20.00 (±3.33)	13.61 (±4.17)

Table E.19. **Accuracy on CelebA for LBC.**

	100%	25%	10%	5%	1%
group-labeled only	77.57 (±1.46)	73.54 (±0.63)	74.86 (±1.25)	76.60 (±1.18)	72.29 (±3.12)
random label		78.19 (±0.14)	78.75 (±0.28)	79.03 (±0.56)	78.89 (±0.14)
psuedo-label		78.06 (±1.11)	77.57 (±0.49)	76.39 (±0.42)	76.25 (±0.14)
CGL		75.49 (±0.21)	76.39 (±0.69)	76.32 (±0.07)	76.81 (±1.39)

Table E.20. Δ_A **on CelebA for LBC.**

	100%	25%	10%	5%	1%
group-labeled only	12.36 (±0.14)	13.75 (±2.08)	15.28 (±3.06)	13.47 (±0.14)	18.19 (±2.92)
random label		21.94 (±0.28)	24.17 (±1.11)	23.61 (±1.11)	25.28 (±0.00)
psuedo-label		12.50 (±1.39)	13.19 (±2.92)	11.11 (±0.28)	10.00 (±0.56)
CGL		12.08 (±0.14)	9.17 (±0.56)	8.47 (±0.42)	8.33 (±0.83)

Table E.21. Δ_M **on CelebA for LBC.**

	100%	25%	10%	5%	1%
group-labeled only	23.61 (±0.83)	26.67 (±3.89)	30.00 (±5.56)	25.83 (±0.83)	35.00 (±5.56)
random label		43.61 (±0.83)	47.22 (±2.78)	45.83 (±3.06)	49.44 (±0.56)
psuedo-label		24.44 (±2.78)	26.11 (±6.11)	21.67 (±1.11)	19.17 (±0.83)
CGL		23.89 (±0.56)	17.50 (±0.83)	16.39 (±1.39)	16.39 (±1.39)

Table E.22. **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.23. 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.24. 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.25. **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.26. 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.27. 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.28. **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.29. 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.30. Δ_M **on COPMAS for LBC.**

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

References

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

[2] 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. 3

[3] 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

[4] 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. 3

[5] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Adv. Neural Inform. Process. Syst.*, 34, 2021. 1

[6] C-Y. Chuang and Y. Mroueh. Fair mixup: Fairness via interpolation. In *Int. Conf. Learn. Represent.*, 2020. 1

[7] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 4

[8] Sanghyuk Chun, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. *Int. Conf. Mach. Learn. Worksh.*, 2019. 5

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

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Int. Conf. Mach. Learn.*, pages 1050–1059. PMLR, 2016. 4

[11] 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. 3

[12] 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 *Int. Conf. Learn. Represent.*, 2018. 3

[13] 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. 3

[14] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Int. Conf. Algo. Learn. Theory*, pages 63–77. Springer, 2005. 4

[15] 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. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 3

[17] 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. 3

[18] 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

[19] 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, 3, 4

[20] 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. 1

[21] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 1

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. 3

[23] 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. 1

[24] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *Int. Conf. Learn. Represent.*, 2019. 4

[25] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *AAAI*, volume 35, pages 2403–2411, 2021. 1

[26] 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, 3

[27] 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. 5

[28] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *Int. Conf. Learn. Represent.*, 2019. 1

[29] 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. 2021. 3

[30] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6902–6911, 2019. 4

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2818–2826, 2016. 5

[32] Robert Torfason, Eirikur Agustsson, Rasmus Rothe, and Radu Timofte. From face images and attributes to attributes. In *Asian Conf. Compu. Vis.*, pages 313–329. Springer, 2016. 1

[33] 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 *Int. Conf. Comput. Vis.*, pages 6023–6032, 2019. 5

[34] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conf. AI, Ethics, and Society*, 2018. 3

[35] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Int. Conf. Learn. Represent.*, 2017. 5

[36] 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