Fairness through partial awareness: Evaluation of the addition of demographic information for bias mitigation methods

Chung Peng Lee [1] **Rachel Hong** [1] **Jamie Morgenstern** [1]

Abstract

Models that effectively mitigate demographic biases have been explored in two common settings: either requiring full access to demographic information in training or omitting demographic information for legal or privacy reasons. Yet in practice, data can be collected in stages or composed of different sources, so data access can be rather flexible, instead of following the two extremes of complete or a lack of access to demographic annotations. We investigate the fairness impact of disclosing more demographic information and find that *demographic-unaware* methods come at a clear cost to certain fairness metrics in comparison to *demographic-aware* methods. We then empirically show the benefits of a *partially-demographic-aware* setup: collecting only a small number of new samples (0.1% of the full set) with demographics for an over-parameterized model can significantly amend this cost (40% gain in worst-group accuracy). Our findings illustrate that simple data collection efforts may effectively close fairness gaps for models trained on data without demographic information.

1 Introduction

Deep learning models have been widely deployed in various real-world applications, yet prior work has found that these models discriminate along the lines of gender or race in high-stakes applications like face recognition and health systems (Buolamwini & Gebru, 2018; Obermeyer et al., 2019). To prevent the disparate harms of model bias, a common approach to address bias has been to promote fairness through *awareness* (Dwork et al., 2012): requiring demographic information during training to ensure that the model's performance toward each demographic group is sufficiently high. Given the sensitive nature of collecting sociodemographic data (Andrus & Villeneuve, 2022; Mason, 2023), the disclosure of this information can impose a trade-off between the privacy of individuals in the data and the fairness of the model. Furthermore, in scenarios like banking services, it can be illegal to collect demographic information to avoid potential biases from decision-makers (Ho & Xiang, 2020).

To abide by privacy regulations and concerns, researchers have proposed *demographic-unaware* techniques (Ashurst & Weller, 2023; Lahoti et al., 2020; Liu et al., 2021; Sohoni et al., 2020), and while these works have shown improved fairness and robustness compared to traditional empirical risk minimization, it remains unclear whether these algorithms put forward are as fair as *demographic-aware* methods. If this gap remains open, *how should we close it*? And in real-world scenarios, *how does additional data help?*

Prior works consider either complete or a lack of access to demographic information in the training, yet, in the real world, industry practitioners often collect more data without consideration of what additional data they need (Holstein et al., 2019), making the type and schema of accessible data flexible and rarely fixed. For instance, a hospital may collect demographic information after realizing potential fairness concerns but would be unable to track demographic information for past patient data. In another case, a company may work with a proprietary face dataset where collecting demographic information for these samples is illegal, while a publicly available dataset may have gender and race annotations. Therefore, this middle-ground *partially-demographic-aware* scenario with access to *some* demographic information can be more realistic than the other two extremes.

We define this middle-ground scenario in two settings: First, we follow the *demographic-scarce regime* (Awasthi et al., 2021) to model cases obtaining unlabeled samples with demographic information from another distribution is easier, or where tracking the demographic information of existing labeled samples is prohibited. Secondly, we use the *partially-annotated group labels* setup in Jung et al. (2022) motivated by the hospital scenario of collecting demographic information from the same distribution but continuing to use prior

[1]University of Washington. Correspondence to: Chung Peng Lee <lee0618@cs.washington.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

data. With two formally defined settings with incomplete demographic information, we are interested in the following research questions:

- *RQ1: How do state-of-the-art methods without demographic information compare to those requiring full demographic information in terms of fairness objectives?*
- *RQ2: How can existing methods that require demographic information easily adapt to incomplete demographic information settings through pseudo-labeling?*
- *RQ3: What is the marginal fairness gain of collecting more samples with demographic information in incomplete demographic information setups?*

To answer these questions, we implement and evaluate multiple bias mitigation methods that are *demographic-aware*, *demographic-unaware*, and *partially-demographic-aware* to bridge between methods with and without demographic information. Our contributions include (1) a controlled and systematic comparison between *demographic-aware* and *demographic-unaware* methods, (2) empirical evidence of the fairness gap between these two scenarios, and (3) a demonstration of the benefits of collecting little additional demographics that substantially improves fairness.

2 Preliminaries

2.1 Setup

Let there be two potentially different distributions P_1 and P_2 over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ where $\mathcal{X}$ denotes the feature space, $\mathcal{A}$ denotes the sensitive attribute space, and $\mathcal{Y}$ denotes the target label space. We let the notion of *group* g be defined by samples with the same tuple (a, y). We denote the set of all groups to be $\mathcal{G} = \mathcal{A} \times \mathcal{Y}$. For clarity, in the rest of the work, we use *demographic group* to emphasize the set of samples with respect to $\mathcal{A}$ only to distinguish from $g \in \mathcal{G}$. Let $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be the loss function, and we consider sensitive attributes a, target labels y and predictions $\hat{y}$ in the binary case, although our setup can be extended to the non-binary case as well.

Below, we define the settings in which demographic information is available to varying degrees.

Demographic-aware. Let dataset $\mathcal{D}_1^{A,Y} = \{(x_i, a_i, y_i)\}_{i=1}^N$ be drawn from P_1.

Demographic-unaware. Let dataset $\mathcal{D}_1^Y = \{(x_i, y_i)\}_{i=1}^N$ be drawn from a distribution P_1^Y over $\mathcal{X} \times \mathcal{Y}$ where P_1^Y is the marginal distribution of P_1.

Partially-demographic-aware.

1. *Demographic-Scarce.* There exist two datasets $\mathcal{D}_1^Y = \{(x_i, y_i)\}_{i=1}^N$ and $\mathcal{D}_2^A = \{(x_i, a_i)\}_{i=1}^M$. $\mathcal{D}_1^Y$ is drawn from P_1^Y over $\mathcal{X} \times \mathcal{Y}$ where P_1^Y is the marginal of a joint distribution P_1. $\mathcal{D}_2^A$ is drawn from P_2^A over $\mathcal{X} \times \mathcal{A}$ where P_2^A is the marginal of P_2.
2. *Partially-Annotated Group Labels.* There exist two datasets $\mathcal{D}_1^{A,Y} = \{(x_i, a_i, y_i)\}_{i=1}^N$ and $\mathcal{D}_1^Y = \{(x_i, y_i)\}_{i=1}^M$ where $\mathcal{D}_1^{A,Y}$ is drawn from P_1 and $\mathcal{D}_1^Y$ is drawn from P_1^Y where P_1^Y over $\mathcal{X} \times \mathcal{Y}$ is the marginal of P_1.

2.2 Fairness metric

We define various fairness metrics in order to evaluate models across demographic groups A.

Equalized Odds (EOD). The notion of equalized odds is to minimize the differences of false positive rate (FPR) and false negative rate (FNR) between the demographic groups with different sensitive attributes (Hardt et al., 2016). A fair classifier that satisfies equalized odds should satisfy Equation 1 $\forall y \in \{0, 1\}$.

$$\mathbb{P}(\hat{y} = 1 | a = 0, y = y) = \mathbb{P}(\hat{y} = 1 | a = 1, y = y) \quad (1)$$

The evaluation metric for EOD is the sum of absolute differences of FPR and FNR between the demographic groups. i.e. EOD $= \left|\mathbb{P}(\hat{y} = 1 | a = 0, y = 0) - \mathbb{P}(\hat{y} = 1 | a = 1, y = 0)\right| + \left|\mathbb{P}(\hat{y} = 0 | a = 0, y = 1) - \mathbb{P}(\hat{y} = 0 | a = 1, y = 1)\right|$

Rawlsian Min-max Fairness (MMF). The idea of Min-max fairness is to maximize the worst-off group performance. Note that the notion of group in MMF is defined by the tuple $g = (a, y)$. The evaluation metric for MMF is simply the worst-group accuracy stated in Equation 2:

$$\min_{(a,y) \in \mathcal{G}} \mathbb{P}(\hat{y} = y | y, a) \quad (2)$$

3 Experiments

3.1 Methods

We evaluate these *demographic-aware* methods:

1. Group DRO (Sagawa et al., 2019) assumes the empirical distribution to be a mix of group distributions and optimizes for the worst-group distribution loss as

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \sup_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim P_g} \left[\ell(\theta, (x, y))\right] \quad (3)$$

2. Last-layer Fairness Finetuning (LastFFT) (Mao et al., 2023) first trains the model with ERM to learn core features for the encoder, and re-trains the last-layer with fairness constraints on a balanced subset.

We use EOD as the fairness constraint in our experiments with the loss defined as

$$\sum_{i=1}^{N} \ell(\theta, (x_i, y_i)) + \alpha(fpr + fnr) \tag{4}$$

where

$$fpr = \left| \frac{\sum_i p_i \cdot (1-y_i) \cdot a_i}{\sum_i a_i} - \frac{\sum_i p_i \cdot (1-y_i) \cdot (1-a_i)}{\sum_i (1-a_i)} \right|$$

$$fnr = \left| \frac{\sum_i (1-p_i) \cdot y_i \cdot a_i}{\sum_i a_i} - \frac{\sum_i (1-p_i) \cdot y_i \cdot (1-a_i)}{\sum_i (1-a_i)} \right|$$

α is a hyper-parameter allowing the regularization strength, and $p_i = \mathbb{P}(\hat{y}_i = 1)$.

3. Balanced ERM simply trains the model on the balanced subset of the full training set detailed in Table 1.

We evaluate these *demographic-unaware* methods:

1. Just Train Twice (JTT) (Liu et al., 2021) first trains the model with ERM and up-weights the samples misclassified from the first stage by a factor of λ_{up} to train a re-weighted ERM again in the second stage.

2. Adversarially Reweighted Learning (ARL) (Lahoti et al., 2020) constructs a learner and adversary pair where the adversary learns the weight for each sample to maximize the classification loss while the learner aims to minimize the loss. Ideally, the adversary assigns higher weights to loss in the minority group (higher loss) to maximize the classification loss as an up-weighting technique.

We evaluate these *partially-demographic-aware* methods incorporating LastFFT via *pseudo-labeling* of demographic annotation:

1. Vanilla Group Labeling (VGL) simply uses the pseudo-labels as ground truth in any task that requires sensitive attribute information $\mathcal{A}$.

2. Confidence-based Group Labeling (CGL) (Jung et al., 2022) searches for a confidence threshold with a validation set to calibrate the confidence rate of the sensitive attribute predictor. For low-confidence samples, the pseudo-labels are drawn from the empirical conditional distribution $\mathbb{P}(\mathcal{A}|\mathcal{Y} = y)$. Otherwise, the predicted pseudo-labels are used as usual.

However, since we do not perceive any significant difference between these two in our empirical results, we only present results with VGL.

3.2 Dataset

We use **CelebA** (Liu et al., 2015), which consists of 200 thousand images of celebrities, as our primary dataset. We let *Male* be the sensitive attribute $\mathcal{A} = 1$ and *Blond Hair* be the target label $\mathcal{Y} = 1$. For the out-of-distribution $\mathcal{D}_2^A$ in the *Demographic-Scarce* scenario, **FairFace** (Karkkainen & Joo, 2021) contains 100 thousand face images balanced across 7 race groups, with gender annotations in the form $\{\text{Male}, \text{Female}\}$.

	Total	Male		Female	
		Blond	Non Blond	Blond	Non Blond
Full CelebA	162,770	1,387	66,874	22,880	71,629
$\mathcal{D}_1$	81,384	693	33,437	11,440	35,814
$\mathcal{D}_2$	81,386	694	33,437	11,440	35,815
Balanced subset	2,772	693	693	693	693
1% of $\mathcal{D}_2^A$	812	6	334	114	358
0.1% of $\mathcal{D}_2^A$	79	0	33	11	35
$\mathcal{D}_{\text{FairFace}}^A$	86744	45986		40758	
0.5% of $\mathcal{D}_{\text{FairFace}}^A$	433	252		181	

Table 1: Frequency by label and gender for CelebA and FairFace dataset splits. The blond hair attribute is not annotated in FairFace.

To study the scenarios that we are interested in, we randomly split the full training set into two subsets, denoted as $\mathcal{D}_1, \mathcal{D}_2$. We let $\mathcal{D}_1$ be the only dataset with available target labels $\mathcal{Y}$ to fix an equal number of target labels for every scenario. In *Demographic-Scarce* setup, we use $(x, y) \in \mathcal{D}_1^Y$ and $(x, a) \in \mathcal{D}_2^A$.[1] To consider the distribution shift of P_2, we use FairFace as another source of demographics denoted as $\mathcal{D}_{\text{FairFace}}^A$. For *demographic-aware* methods, the training is done completely on $\mathcal{D}_1^{A,Y}$ with the ground-truth demographic information $\mathcal{A}$ of all samples in $\mathcal{D}_1$. For *demographic-unaware* methods, the training is done on $\mathcal{D}_1^Y$ without access to the ground-truth $\mathcal{A}$. As for *Partially-Annotated Group Labels* setup, we disclose the ground-truth demographic information for some samples in $\mathcal{D}_1$ such that $\mathcal{D}_1$ is partially group-annotated with some $(x, a, y) \in \mathcal{D}_1^{A,Y}$ and $(x, y) \in \mathcal{D}_1^Y$. Table 1 summarizes the number of samples in each setup for each group.

Finally, access to a full demographic information validation set is assumed throughout our experiments in any setup. This can plausibly be implemented in practice through a trusted third-party auditor without disclosing individual demographic information to the model producers.

3.3 Model

For all prediction models, we use ResNet50 (He et al., 2016) from *torchvision* with ImageNet pre-trained weights initial-

[1]The superscript indicates the accessibility of $\mathcal{A}$ and/or $\mathcal{Y}$ for any set $\mathcal{D}$.

ization as the backbone. For the pseudo-labeling task, we use both ResNet50 and a fully connected Linear model as sensitive attribute predictors. The Linear sensitive attribute predictor serves as a weaker model to show the impact of collecting more in-distribution demographic information.

Throughout the experiments, we observe overfitting to fairness metrics for all methods through a validation set with full demographic information. Therefore, we select the final model with the lowest validation fairness violation which does not overfit depending on the specific fairness constraint in training.

4 Main Results

4.1 Fairness gap between *demographic-aware* and *demographic-unaware*

We observe that *demographic-unaware* methods do significantly worse than *demographic-aware* methods for both worst-group accuracy and equalized odds. From Figure 1 and Table 2, the equalized odds gap remains large, while the worst-group accuracy gap is less significant and partially closed by JTT at the cost of overall accuracy. From observing the fairness-accuracy tradeoff differences between *demographic-aware* and *demographic-unaware* methods, we study *how more demographically-annotated data can bridge this gap* in the next section.

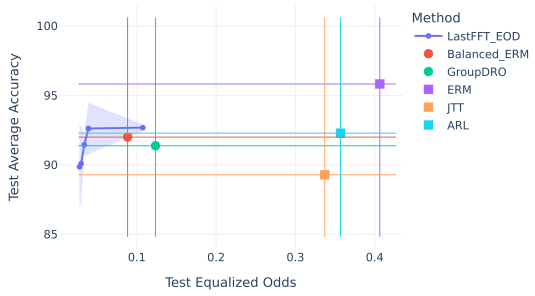


Figure 1: Equalized odds and accuracy tradeoff on the standard test set for demographic-aware and unaware methods. The more upper-left a method is, the better it is, with lower equalized odds violation and higher overall accuracy. LastFFT's curve comes from a sweep of its different hyperparameters α's and taking the Pareto set of the results.

4.2 Marginal fairness return of new samples with demographic information

The ResNet50 sensitive attribute predictor converges to a high accuracy rate with just 0.1% of $\mathcal{D}_2^A$. However, for the Linear sensitive attribute predictor, with the increase of samples with demographic information, Figure 2a shows a clear left-shift of the fairness-accuracy tradeoff curve. Moreover, comparing Figures 2a and 2b, we see clearer improvement in the *Partially-Annotated Group Labels* setup but a less consistent one in the *Demographic-Scarce* setup. This suggests the potential bottleneck of pseudo-labels: collecting more samples (x, a) from a distinct dataset may not provide a more accurate estimate of $\mathcal{A}$ in existing samples (x, y), in spite of being drawn from the same distribution. On the other hand, annotating $\mathcal{A}$ for existing samples would always provide a closer-to-ground-truth estimate but can be more costly. This can be further supported by the out-of-distribution case in Figure 2c where collecting more demographics from FairFace shows initial improvement but reaches a bottleneck.

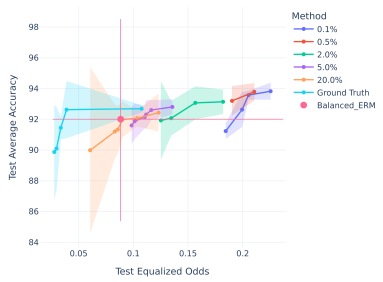
	Avg Acc	Worst-Group Acc	EOD	Data Scenario
Group DRO	$91.40_{\pm 0.28}$	$78.88_{\pm 1.17}$	$0.12_{\pm 0.00}$	$(x,a,y) \in \mathcal{D}_1^{A,Y}$
LastFFT$_{\alpha=2.5}$	$91.49_{\pm 1.85}$	$86.79_{\pm 1.38}$	$0.05_{\pm 0.03}$	$(x,a,y) \in \mathcal{D}_1^{A,Y}$
Balanced ERM	$91.87_{\pm 0.47}$	$82.56_{\pm 1.34}$	$0.08_{\pm 0.02}$	$(x,a,y) \in \mathcal{D}_1^{A,Y}$
LastFFT$_{\alpha=10}$	$92.14_{\pm 1.57}$	$84.87_{\pm 2.96}$	$0.12_{\pm 0.03}$	$(x,a,y) \in 0.1\%\ \mathcal{D}_1^{A,Y}$ and $(x,y) \in 99.9\%\ \mathcal{D}_1^{Y}$
LastFFT$_{\alpha=2.5}$	$92.60_{\pm 0.91}$	$84.45_{\pm 3.22}$	$0.14_{\pm 0.03}$	$(x,a) \in 0.1\%\ \mathcal{D}_2^{A}$ and $(x,y) \in \mathcal{D}_1^{Y}$
LastFFT$_{\alpha=20}$	$90.78_{\pm 1.98}$	$85.15_{\pm 3.16}$	$0.15_{\pm 0.03}$	$(x,a) \in 0.5\%\ \mathcal{D}_{\text{FairFace}}^{A}$ and $(x,y) \in \mathcal{D}_1^{Y}$
JTT	$89.37_{\pm 0.38}$	$73.89_{\pm 1.18}$	$0.33_{\pm 0.01}$	$(x,y) \in \mathcal{D}_1^{Y}$
ARL	$91.31_{\pm 3.61}$	$60.0_{\pm 15.27}$	$0.36_{\pm 0.11}$	$(x,y) \in \mathcal{D}_1^{Y}$
ERM	$95.81_{\pm 0.13}$	$44.00_{\pm 6.40}$	$0.44_{\pm 0.05}$	$(x,y) \in \mathcal{D}_1^{Y}$

Table 2: Various fairness metrics of individual methods across each data scenario with a fixed number of labeled samples on CelebA[2]. We select a few LastFFT points on the fairness-accuracy tradeoff curve to demonstrate how little additional demographics can provide significant improvement. The pseudo-labeling is done by ResNet50 sensitive attribute predictor with VGL.

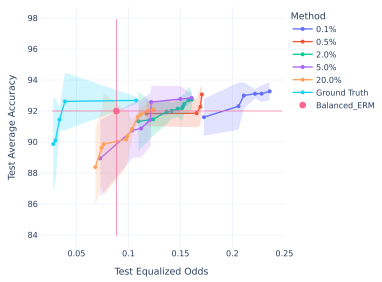
4.3 Significant fairness improvement with little demographics

In the *Demographic-Scarce* setup with a ResNet50 sensitive attribute predictor, we observe that by adding only 79 additional **in-distribution** unlabeled samples with demographic information (0.1% of $\mathcal{D}_2^A$), the *partially-demographic-aware* LastFFT method with pseudo-labeling achieves significant improvement compared to any *demographic-unaware* method, as shown in Table 2. The out-of-distribution case further provides additional promise as obtaining demographics from out-of-distribution data may be easier and less expensive. One configuration of LastFFT presented in Table 2 substantially outperforms *demographic-unaware* methods in any fairness metric with only 433 ad-

[2]The results for JTT and GroupDRO differ from the original work because we only use half of the full CelebA training set.



(a) *Partially-Annotated Group Labels* setup: equalized odds and accuracy tradeoff with additional demographic annotations to train Linear sensitive attribute predictor.



(b) *Demographic-Scarce* setup: equalized odds and accuracy tradeoff with additional demographic annotations from **CelebA** to train a Linear sensitive attribute predictor.

(c) *Demographic-Scarce* setup with distribution shift: equalized odds and accuracy tradeoff with additional demographic annotations from **FairFace** to train a ResNet50 sensitive attribute predictor.

Figure 2: Fairness-accuracy plots for LastFFT configurations in the *partially-demographic-aware* regime. The Ground Truth baseline refers to LastFFT trained on complete $\mathcal{D}_1^{A,Y}$ from Figure 1. Balanced ERM is a *demographic-aware* method as defined in Section 3.1. *Demographic-unaware* baselines are located to the right of each subplot and are not included for visualization purposes.

ditional **out-of-distribution** unlabeled samples with demographics (0.5% of $\mathcal{D}^A_{\text{FairFace}}$).

5 Discussion

5.1 Implications of demographic imputation

We acknowledge the societal concerns of inferring demographic information in certain situations. Using demographic proxies can be prohibited according to "anti-classification" principles of U.S. anti-discrimination law (Ho & Xiang, 2020). Prior work, however, has demonstrated its necessity to mitigate (much less evaluate) performance disparities when demographic information is not recorded (Cheng et al., 2023; Rieke et al., 2022). This tension reflects ongoing legal debates on the differences between disparate impact and disparate treatment of individuals (Barocas & Selbst, 2016; Siegel, 2003). In our work, we aim to understand the most intuitive way of bridging the gap between *demographic-aware* and *demographic-unaware* methods through pseudo-labeling. **We do not encourage the explicit use of pseudo-labels in training without consideration of potential harms as a result of group misclassification.** The robust results we observe in Balanced ERM in Table 2 with a small subset of samples inspire the potential use of pseudo-labeling to construct an approximately balanced ERM without directly using demographics.

5.2 Limitations and future work

As Gulrajani & Lopez-Paz (2020) highlight, methods may not generalize to other datasets, so we aim to follow our same framework for additional datasets to understand whether the conclusions we draw from this work are consistent. We also examine several well-known fairness methods but leave evaluations of other methods like contrastive-learning-based algorithms to future work. While prior works (Nam et al., 2022; Jung et al., 2022) in the *partially-demographic-aware* realm focus on pseudo-labeling, we hope to incorporate methods that inherently address incomplete demographic information without pseudo-labeling.

6 Conclusion

In this work, we first ask whether it suffices to use state-of-the-art *demographic-unaware* methods to preserve privacy in comparison to *demographic-aware* ones. After finding empirical evidence of the fairness gap between these two categories, we are motivated by the flexibility of data collection in the real world and investigate how to close this gap through access to additional demographic information. Specifically, how much demographic information do we need to *bridge the gap*? For the CelebA task, we find that the addition of 79 in-distribution samples or 433 out-of-distribution samples with demographics is sufficient to gain significant improvement along various fairness metrics compared to *demographic-unaware* methods. As a result, our work questions the assumptions that a large amount of demographic information is necessary to improve group fairness measures. Finally, we address the ethical concerns in this work regarding the use of pseudo-labels in Section 5.1. The goal of this work is to encourage more development of *partially-demographic-aware* methods without the use of pseudo-labels and measure how far we have to go.

Impact Statement

This work is directly motivated by preventing societal harms of machine learning models, by considering realistic scenarios in which demographic information may not always be feasible in the data collection step. We note that some of our methods incorporate training a sensitive attribute predictor and that this artifact immediately imposes potential risks of misuse by other users. Thus, we do not store any pseudo-labels for any sample, and we choose to not release any trained sensitive attribute predictors. At the same time, studying the use of pseudo-labels in cases when demographic information is unobserved allows us to understand how to build safer, fairer machine learning models in data-restricted settings.

References

Andrus, M. and Villeneuve, S. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1709–1721, 2022.

Ashurst, C. and Weller, A. Fairness without demographic data: A survey of approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–12, 2023.

Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., and Wang, X. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 206–214, 2021.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Cheng, L., Gallegos, I. O., Ouyang, D., Goldin, J., and Ho, D. How redundant are redundant encodings? blindness in the wild and racial disparity when race is unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 667–686, 2023.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ho, D. E. and Xiang, A. Affirmative algorithms: The legal grounds for fairness as awareness. *U. Chi. L. Rev. Online*, pp. 134, 2020.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–16, 2019.

Jung, S., Chun, S., and Moon, T. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10348–10357, 2022.

Karkkainen, K. and Joo, J. 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*, pp. 1548–1558, 2021.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Mao, Y., Deng, Z., Yao, H., Ye, T., Kawaguchi, K., and Zou, J. Last-layer fairness fine-tuning is simple and effective for neural networks. *arXiv preprint arXiv:2304.03935*, 2023.

Mason, P. L. Striking the balance: Approaches to racial equitable data collection that protect privacy in health. 2023.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Rieke, A., Southerland, V., Svirsky, D., and Hsu, M. Imperfect inferences: A practical assessment. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 767–777, 2022.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv*, abs/1911.08731, 2019. URL https://api.semanticscholar.org/CorpusID:208176471.

Siegel, R. B. Equality talk: Antisubordination and anticlassification values in constitutional struggles over brown. *Harv. L. Rev.*, 117:1470, 2003.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

A Additional Results

A.1 Sensitive Attribute Predictor

Figure A.1 provides results on how each sensitive attribute predictor performs for predicting gender on the standard test set of CelebA. For the in-distribution *Demographic-Scarce* setup versus *Partially-Annotated Group Labels* setup, there is no explicit difference in terms of test accuracy. For the out-of-distribution *Demographic-Scarce* setup, we see a significant drop from the in-distribution counterpart with the same model ResNet50.

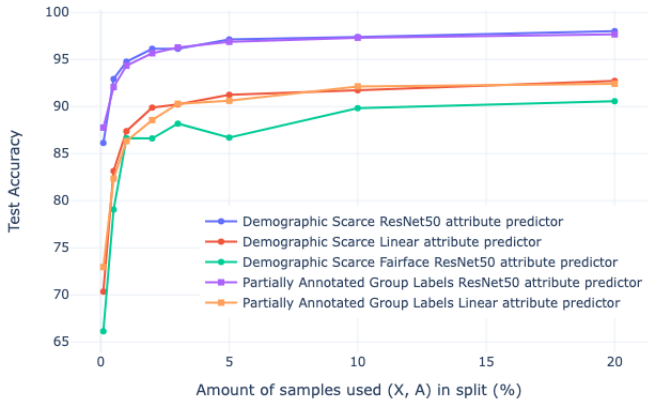
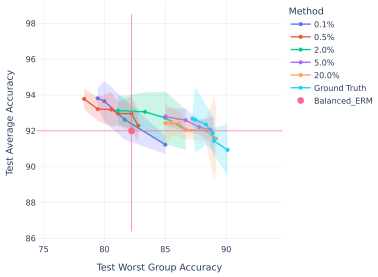


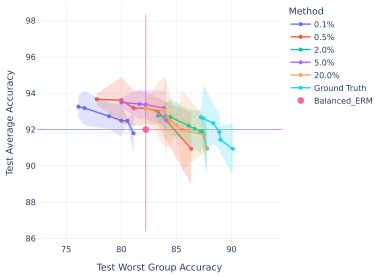
Figure A.1: Test accuracy of different sensitive attribute predictors on CelebA's standard test set

A.2 Marginal Min-max Fairness and Accuracy Tradeoff Return of Collecting Demographic Information



(a) *Partially-Annotated Group Labels* setup min-max fairness and accuracy tradeoff with increasing demographics collected and trained on for a Linear sensitive attribute predictor

(b) *Demographic-Scarce* setup min-max fairness and accuracy tradeoff with increasing demographics collected from **CelebA** and trained on for a Linear sensitive attribute predictor



(c) *Demographic-Scarce* setup min-max fairness and accuracy tradeoff with increasing demographics collected from **FairFace** and trained on for a ResNet50 sensitive attribute predictor

Figure A.2

B Training Details

B.1 Setup

For hyper-parameters search, we use a subset of the full data and less number of epochs to find a good configuration of hyper-parameters by a grid search. Then, we use the full set depending on the setup and methods with the best configuration. The compute resources include 2 NVIDIA A100 GPUs.

B.2 Empirical Risk Minimization (Blond Hair)

B.2.1 Hyper-parameters Search

Hyper-parameter	Search range
learning rate	[1e-3, 1e-4]
ℓ_2 regularization strength	[1e-2, 1e-3, 1e-4]
data subset	30%
# of epochs	10
batch size	128

Table B.1: Hyper-parameter grid search range for ERM

B.2.2 Hyper-parameters and Configuration

Learning rate	ℓ_2 regularization strength	Optimizer	# of Epochs
1e-4	1e-3	AdamW	10

Table B.2: Hyper-parameters and configuration for ERM

B.3 Empirical Risk Minimization (Male)

B.3.1 Hyper-parameters Search

Hyper-parameter	Search range
learning rate	[1e-3, 1e-4]
ℓ_2 regularization strength	[1e-2, 1e-3, 1e-4]
data subset	30%
# of epoch	10
batch size	128

Table B.3: Hyper-parameter grid search range for ERM

B.3.2 Hyper-parameters and Configuration

Learning rate	ℓ_2 regularization strength	Optimizer	# of Epochs
1e-4	1e-3	AdamW	10

Table B.4: Hyper-parameters and configuration for ERM

B.4 Group DRO

B.4.1 Hyper-parameters and Configuration

For Group DRO, we use the same hyper-parameter configuration as Sagawa et al. (2019) without more search.

Learning rate	ℓ_2 regularization strength	η_q	Optimizer	# of Epoch
1e-4	0.1	0.01	SGD	50

Table B.5: Hyper-parameters and configuration for Group DRO

B.5 Last-layer Fairness Fine-tuning (Equalized Odds)

B.5.1 Hyper-parameters Search

In Mao et al. (2023), they use $\alpha = 10.0$ as the best hyper-parameter for α. However, since we are sweeping through different α's to obtain fairness-accuracy tradeoff curves, we use $\alpha = 10.0$ to search for other hyper-parameters, and then we sweep through the range of α's with the best configuration.

Hyper-parameter	Search range
learning rate	[1e-3, 1e-4, 1e-5]
ℓ_2 regularization strength	[1e-1, 1e-2, 1e-3, 1e-4]
α	10.0
data subset	balanced subset
# of epochs	50
# of pre-training epochs	1
batch size	128

Table B.6: Hyper-parameter grid search range for Last-layer Fairness Fine-tuning w.r.t. Equalized Odds

B.5.2 Hyper-parameters and Configuration

Learning rate	ℓ_2 regularization strength	α	Optimizer	# of Epochs
1e-3	1e-4	[0.0, 0.1, 0.2, 0.5, 1.0, 5.0, 10.0]	AdamW	50

Table B.7: Hyper-parameters and configuration for Last-layer Fairness Fine-tuning w.r.t. Equalized Odds

B.5.3 Construction of Balanced Subset

In the second stage of LastFFT, a balanced subset is required for fine-tuning. However, in *partially-demographic-aware* setup, a true balanced subset is not accessible. Therefore, we use pseudo-labels $\hat{a}$ to construct an *approximately balanced subset*. The construction is by taking k samples from each proxy group $\hat{g} = (\hat{a}, y)$ where k is the size of the smallest proxy group such that the resulting set consisting of $4k$ samples in total.

B.6 Just Train Twice

B.6.1 Hyper-parameters Search

We do a hyper-parameter search for λ_{up} because our setup consists of half of the samples with target labels. We find that the optimal λ_{up} in our setup is exactly half of the optimal λ_{up} found in Liu et al. (2021).

Hyper-parameter	Search range
λ_{up}	[10, 25, 50]
learning rate	1e-5
ℓ_2 regularization strength	1e-1
data subset	100%
# of epochs	30
# of pre-training epochs	1
batch size	128
gradient accumulation step	4

Table B.8: Hyper-parameter grid search range for Just Train Twice

B.6.2 HYPER-PARAMETERS AND CONFIGURATION

Learning rate	ℓ_2 regularization strength	λ_{up}	Optimizer	Epoch
1e-5	1e-1	25	SGD	50

Table B.9: Hyper-parameters and configuration for Just Train Twice

B.7 Adversarially Reweighted Learning (ARL)

B.7.1 HYPER-PARAMETERS SEARCH

We include the adversary architecture in the hyper-parameter search. We use the gradient accumulation technique to average the gradients of a larger batch after observing the instability of training ARL.

Hyper-parameter	Search range
learner learning rate	[1e-2, 1e-3, 1e-4]
adversary learning rate	[1e-2, 1e-3, 1e-4]
ℓ_2 regularization strength	[1e-1, 1e-2]
adversary architecture	[shared encoder linear, separate encoder, linear]
gradient accumulation step	16
data subset	50%
# of epochs	30
# of pre-training epochs	1
learning rate (pre-training)	1e-4
ℓ_2 regularization strength (pre-training)	1e-1
batch size	64

Table B.10: Hyper-parameter grid search range for Adversarially Reweighted Learning

B.7.2 HYPER-PARAMETERS AND CONFIGURATION

Learning rate	Adversary learning rate	ℓ_2 regularization strength	Adversary	Optimizer	Epoch
1e-2	1e-2	shared encoder linear	1e-1	AdamW	50

Table B.11: Hyper-parameters and configuration for Adversarially Reweighted Learning

B.7.3 LEARNER-ADVERSARY ARCHITECTURE

While Lahoti et al. (2020) propose ARL as a flexible framework, the implementation for deep neural networks remains unclear. Let Z be the embedding after the feature extractor, h_{θ_1} be the feature extractor of ResNet50, and c_{θ_2} be the classification layer. Let f_ϕ be the adversary in the ARL framework. We experiment with three learner-adversary architectures. *Shared encoder linear* refers to the one shown in Figure B.1 where the adversary is a linear layer taking the concatenation of embeddings and target label as the input. We find this to be the best-performing learner-adversary architecture. For *separate encoder* architecture, the learner and adversary each have their encoders where the adversary takes the image as the input and the gradients of adversary weights update the adversary encoder as well. Finally, *linear* refers to a single fully connected layer on the raw image of size $3 \times 224 \times 224$ flattened.

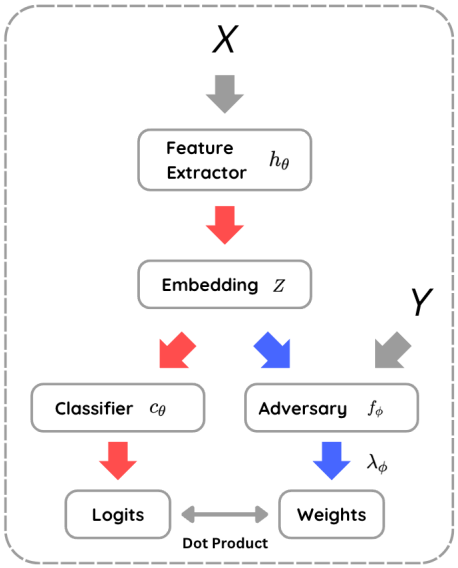


Figure B.1: ARL shared encoder architecture