Scaling Bias Mitigation with Multiple Fairness Tasks and Multiple Protected Attributes

Eric Zhao *University of California, Berkeley* *eric.zh@berkeley.edu*

De-An Huang *Nvidia Research* *deahuang@nvidia.com*

Hao Liu *California Institute of Technology* *hliu3@caltech.edu*

Zhiding Yu *Nvidia Research* *zhidingy@nvidia.com*

Anqi Liu *Johns Hopkins University* *aliu@cs.jhu.edu*

Olga Russakovsky *Princeton University* *olgarus@cs.princeton.edu*

Anima Anandkumar *Nvidia Research* *aanandkumar@nvidia.com*

Abstract

Bias mitigation methods are commonly evaluated with a single fairness task, which aims to reduce performance disparity with respect to a single protected attribute (e.g., gender) while maintaining predictive performance for target labels (e.g., is-cooking). In this work, we question whether this mode of evaluation provides reliable insights into the effectiveness of bias mitigation methods. First, there are multiple protected attributes in real-world applications, such as skin color, gender and age. Second, we find that the results of these studies vary greatly depending on the choice of fairness task for evaluation. We address these shortcomings by first evaluating bias mitigation methods on the CelebA dataset on 54 different fairness tasks, which involve various selections and intersections of multiple protected attributes. Our thorough analysis shows that simple importance weighting is still a consistently competitive method for bias mitigation. We then extend our analysis to ImageNet's People Subtree, which poses qualitatively different real-world challenges than CelebA: having hundreds of protected groups while fewer than 10% of the training dataset has protected attribute labels. We find that strategies to reduce model complexity are important in this scenario. We show that leveraging these insights can reduce the bias amplification of empirical risk minimization by 28% on ImageNet's People Subtree.

1 Introduction

There is a significant potential for harm when the predictive properties of machine learning models vary across different demographic populations, e.g., protected groups. This is indeed the case for many real-world applications including facial analysis (Buolamwini & Gebru, 2018), ad delivery (Sweeney, 2013) and search engines (Noble, 2018). Algorithmic bias is particularly challenging to address in deep learning systems, as obtaining formal guarantees is impractical and commonly-used algorithmic fairness techniques do not scale.

Recent works seek means of harm reduction by proposing bias mitigation methods for deep learning (Wang et al., 2020; Sagawa et al., 2020; Liu et al., 2021; Zhao et al., 2017; Edwards & Storkey, 2015; Arjovsky et al., 2020; Pezeshki et al., 2020). These methods aim to empirically reduce quantitative measures of algorithmic bias without significantly harming overall performance. A commonly used dataset for empirically validating these methods is CelebA (Liu et al., 2015), which labels celebrity images with 40 attributes such as hair color. Bias mitigation methods are often evaluated on a single *fairness task* on CelebA, which aims to maintain a

computer vision model's predictive performance on a *target label* while reducing bias metrics with respect to a *protected attribute* (e.g., gender). The target label and protected attribute for the fairness task are selected by the experiment designer from CelebA's set of 40 available attributes.

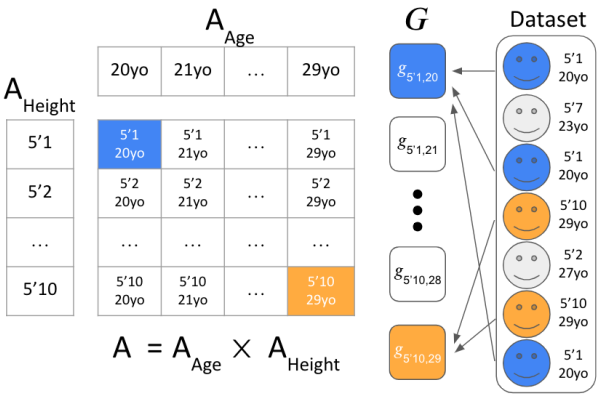
We question whether this mode of evaluation truly provides reliable insights into the effectiveness of bias mitigation methods. First, there are many protected attributes in real-world applications, such as skin color, gender and age; bias mitigation methods need to protect not only individual attributes, but also the *intersections* of protected attributes (Crenshaw, 1989). Second, it is unclear how much the findings depend on the particular fairness task selected for evaluation. To this end, we describe two novel sets of experiments: (1) we evaluate bias mitigation methods on 54 CelebA fairness tasks by sweeping over different choices of protected attributes and different intersections of these protected attributes; and (2) we adapt bias mitigation methods for *label-scarce* settings and apply them to a real-world algorithmic bias problem on the ImageNet dataset.

Summary of Results: In Section 4, we study how sensitive the results of CelebA bias mitigation experiments are to their choices of protected attributes. We survey a diverse selection of bias mitigation methods and evaluate them on many different choices of protected groups, including some intersectional groups that we induce from the intersections of multiple protected attributes. Our analysis finds that the relative performances of bias mitigation methods vary greatly depending on which protected groups the experiment designer selects; this results points to the unreliability of CelebA experiments in prior works that test their methods on only one choice of protected groups. Our comprehensive experiments find that, in fact, simple importance weighting regularly outperforms all other bias mitigation methods. We also identify several—to the best of our knowledge—previously unreported trends, including the sensitivity of reweighting-based methods to noisy protected attribute labels and the non-monotonicity of popular bias metrics in the transition from single-attribute to intersectional multi-attribute tasks.

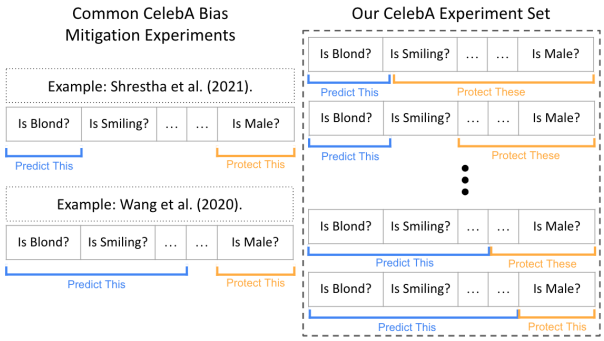
In Section 5, we extend our critical analysis of bias mitigation methods to the ImageNet dataset (Deng et al., 2009), where Yang et al. (2020) recently labeled the People Subtree with protected attributes such as age and gender. Whereas designations of "target" and "protected" attributes on CelebA are artificially specified by an experiment designer, ImageNet poses a well-defined real-world algorithmic fairness challenge, lending a more realistic test environment for bias mitigation. ImageNet also poses qualitatively different challenges than CelebA: the underlying learning task is more challenging (hundreds of target labels/classes), there are more protected groups to address (196), and protected attribute labels are scarce (less than 10% of images labeled). We find that these factors limit the effectiveness of existing bias mitigation methods, and thus propose techniques for "adapting" these methods to ImageNet: *knowledge distilation* and *attribute decomposition*. Our experiments provide the first evaluation of bias mitigation methods on the ImageNet dataset and singles out two methods as being particularly effective after being properly "adapted": importance weighted ERM and Wang et al. (2020)'s "domain independence" method.

2 Related Work

Bias Mitigation in Deep Learning Prior works have highlighted concerns of algorithmic bias in a number of real-world deep learning applications, including commercial image classifiers (Buolamwini & Gebru, 2018) and natural language processors (Alvi et al., 2018; Bolukbasi et al., 2016; Garg et al., 2018). However, classical algorithmic fairness techniques are of limited applicability for these deep learning settings, as they typically either assume simple models (e.g. linear regression) or appeal to costly procedures like no-regret dynamics (Agarwal et al., 2018; Saerens et al., 2002; Dwork et al., 2012; Zhang et al., 2018). Recent literature have instead sought bias mitigation methods designed for deep learning applications (Edwards & Storkey, 2015; Ramaswamy et al., 2021; Ryu et al., 2018; Wang et al., 2020; Zhao et al., 2017). They borrow on techniques from related fields including robust optimization (Adragna et al., 2020; Liu et al., 2021; Sagawa et al., 2020), causal inference (Arjovsky et al., 2020; Creager et al., 2021; Kusner et al., 2017; Madras et al., 2019), and representation learning (Pezeshki et al., 2020). As obtaining theoretical guarantees for deep learning bias mitigation methods is impractical, prior works have resorted to various choices of summary statistics for quantifying algorithmic bias; these metrics include worst-case accuracy (weighted on the worst-off group), mean accuracy (balanced over protected groups), bias amplification scores (Zhao et al., 2017), and



A graphic illustration of "protected groups" and "protected attributes" (see Section 3). The protected attributes are Age and Height: Attrs = {Age, Height}. The Age attribute has 10 possible values A_{Age}; the Height attribute also has 10 values A_{Height}. A is the set of all Height-Age pairs. The protected groups, G, partition the dataset based on each datapoint's protected attributes. Each group $g \in G$ is associated with an intersectional identity in A, for instance the blue entries in G and A.



A graphic illustration of CelebA bias mitigation experiment designs used in prior works to validate their proposed methods (left) and the experiment designs we propose in this work (right). Each experiment design designates some CelebA attributes (e.g., "Is Blond?") to predict and an attribute (e.g., "Is Male?") to mitigate bias against. We find these design choices significantly influence experiment results. Our experiments sweep over many experiment designs including "intersectional" settings with multiple protected attributes.

intersectional bias scores (Foulds et al., 2020). Other works have proposed evaluation mechanisms that explicitly generate counterfactuals to measure bias but require significant computational or human resources (Balakrishnan et al., 2020; Denton et al., 2020).

Intersectional Multi-Attribute Fairness One of our goals is extending prior empirical analyses of bias mitigation methods to intersectional settings (Wang et al., 2020; Shrestha et al., 2021; Liu et al., 2021; Sagawa et al., 2020). Prior works have proposed methods for specifically addressing intersectional fairness concerns at classical learning settings (Kang et al., 2021; Kearns et al., 2018; Foulds et al., 2020; Makar et al., 2021; Wang et al., 2022). In this work, we take a step further and perform a thorough empirical analysis on bias mitigation for large-scale computer vision tasks. While our work focuses on empirical analyses and largely abstracts away from social considerations, other works have directly connected the technical constraints of algorithmic intersectional bias mitigation to social implications (see, e.g., Wang et al. (2022) and Kong (2022)).

Label-Scarce Bias Mitigation In many settings, it may not be practical to label the protected attributes of all datapoints; this is indeed the case for ImageNet, where less than 10% of the People Subtree has been annotated with protected attributes (Deng et al., 2009; Yang et al., 2020). The problem of bias mitigation in settings where protected attribute information is scarce has been studied by Dai & Wang (2021) for graph neural networks and Ho et al. (2020) for adversarial learning. Other works have studied entirely unsupervised bias mitigation (Chen et al., 2019; Hashimoto et al., 2018) and the use of proxy labels for protected attributes (Kallus et al., 2020; Awasthi et al., 2021).

3 Preliminaries

Problem Formulation Formally, we are interested in the task of learning a classification model $h : \mathcal{X} \to \mathcal{Y}$ that maps from a feature space $\mathcal{X}$ to a label space $\mathcal{Y}$. For instance, if we desire h to predict whether a photographed individual has blond hair, $\mathcal{X}$ may be the space of 256x256 grayscale images and $\mathcal{Y}$ the binary set $\{\text{yes}, \text{no}\}$. We are also interested in our model h being "fair" with respect to a set of *protected groups* $\mathcal{G}$. Continuing our example, if $\mathcal{G}$ is the set of all age groups, we may desire our model's predictions of whether a photographed individual has blond hair to be independent of the individual's age. Precise notions of algorithmic fairness have been debated extensively in literature, but the most commonly accepted include demographic parity (Feldman et al., 2015) and equalized odds (Hardt et al., 2016).

Protected Attributes We define protected groups as arising from a set of *protected attributes*, Attrs. We denote the set of potential values of a protected attribute $a \in \text{Attrs}$ as $\mathcal{A}_a$, and use $\mathcal{A} := \prod_{a \in \text{Attrs}} \mathcal{A}_a$ to denote the set of all possible intersections of protected attributes. We accordingly define a protected group g as individuals who share protected attributes and the set of all protected groups $\mathcal{G}$ as isomorphic to $\mathcal{A}$. In our running example, $\mathcal{G}$ arises from choosing "age" as a single protected attribute where Attrs = [age], $\mathcal{A} = \mathcal{A}_{\text{age}} = \{\text{Children}, \text{Young Adults}, \dots\}$.

When $|\text{Attrs}| = 1$, i.e. we are only concerned with a single protected attribute, we say that our set of protected groups $\mathcal{G}$ is not intersectional. When $k := |\text{Attrs}| > 1$, we say that our set of protected groups $\mathcal{G}$ is k-intersectional, as they arise from the intersections of k protected attributes. In these cases, it is important that our learned model h satisfies *intersectional fairness* (Foulds et al., 2020). That is, we not only desire h to treat similarly older vs. younger individuals and longer vs. shorter haired individuals, but also their intersections: e.g. older longer-haired individuals and younger shorter-haired individuals.

Fair Learning Datasets In the datasets we use to evaluate bias mitigation methods, each datapoint consists of the tuple $(x, y, g) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{G}$. That is, datapoints not only provide features x and labels y for supervised learning, but also which protected groups g they correspond to. We also encounter *label-scarce* datasets where x, y are available for all datapoints, but protected group labels g are only available for a small subset of datapoints. We will use the adjectives "labeled"/"unlabeled" to refer to whether g is available—not whether y is available. "Unlabeled bias mitigation methods" are those that only use x, y and not g.

3.1 Measures of Algorithmic Bias

Given that obtaining theoretical guarantees is impractical in deep learning settings, an important aspect of evaluating bias mitigation algorithms is identifying quantitative metrics of their efficacy. However, it is difficult to quantify algorithmic bias with a single real-valued statistic, as traditional notions of algorithmic fairness like demographic parity and equalized odds are matrix-valued constraints. Prior works have instead identified useful surrogate metrics that offer more meaningful and less noisy feedback than, for instance, naively averaging over the matrix of demographic parity violations (Wang et al., 2020; Shrestha et al., 2021; Liu et al., 2021). We similarly adopt the use of these surrogate metrics for our analyses.

Importance-weighted accuracy (or "reweighted accuracy") is an importance weighted accuracy metric, where importance weights are selected so that each protected group has equal weight:

$$\text{Acc}_U(h) = \sum_{g \in \mathcal{G}} \frac{1}{\Pr(G = g)} \Pr(Y = h(X) \mid G = g) \tag{1}$$

In multi-label settings, we equivalently define a notion of reweighted Mean Average Precision (mAP). High reweighted accuracy or reweighted mAP is a strong indicator of good predictive performance and a weak indicator that the model does not significantly overfit to heavily represented groups.

Bias amplification measures the difference between a group's ground-truth representation for a label class versus the group's predicted representation (Zhao et al., 2017):

$$s_y := \Pr(G = g \mid h(X) = y) - \Pr(G = g \mid Y = y) \text{ where } g := \text{argmax}_{g \in \mathcal{G}} \Pr(G = g \mid h(X) = y). \tag{2}$$

This score does not signal predictive performance, but a positive score strongly indicates that the model may amplify prediction biases in the data. Thus, we generally seek a bias amplification score close to or below zero, though non-positive scores are not proof for the absence of bias. This score is vector-valued (defined per class y), so when target labels are non-binary $|\mathcal{Y}| > 2$, we average the score vector: $s = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} s_y$. Bias amplification have been widely adopted by prior works like Wang et al. (2020), but has been found to be limited in what it captures (Wang & Russakovsky, 2021).

The intersectional bias score (Foulds et al., 2020) summarizes demographic parity violations.

$$\epsilon_y := \max_{g_1, g_2 \in \mathcal{G}} \text{argmin}_\epsilon \text{ s.t. } \exp(-\epsilon) \leq \frac{\Pr(h(X) = y \mid G = g_1)}{\Pr(h(X) = y \mid G = g_2)} \leq \exp(\epsilon)$$

Table 1: List of the bias mitigation algorithms evaluated in Sections 4 & 5.

Name	Abbreviation	Needs Attribute Labels	Reweighting	Adversarial
Empirical Risk Minimization (ERM)	U-ERM	No	No	No
Importance Weighted Empirical Risk Minimization	W-ERM	Yes	Yes	No
Group D. Robust Optim. (Sagawa et al., 2020)	WA-GDRO	Yes	Yes	Yes
Adversarial Censoring (Edwards & Storkey, 2015)	A-Cens	Yes	No	Yes
Invariant Risk Minimization (Arjovsky et al., 2020)	IRM	Yes	No	No
Domain Independence (Wang et al., 2020)	Ind	Yes	No	No
Domain Discriminative (Dwork et al., 2012)	Disc	Yes	No	No
Spectral Decoupling (Pezeshki et al., 2020)	U-SD	No	No	No
Just Train Twice (Liu et al., 2021)	UWA-JTT	No	Yes	Yes

Similar to bias amplification, this score does not signal predictive performance. A high intersectional bias score indicates significant violations of the demographic parity constraint. These scores are also vector-valued and defined per class y, so we again average the score vectors: $\epsilon = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \epsilon_y$.

3.2 Bias Mitigation Methods

Our experiments study a diverse and representative set of bias mitigation methods. We list these in Table 1 and categorize them along three axes: (1) do the methods need protected group labels, (2) do they involve an adversary in their algorithm, and (3) is their mechanism of action the manipulation of importance weights.

Baselines Our studies include, as baselines, *empirical risk minimization* and *importance weighted empirical risk minimization*. We use "empirical risk minimization" to refer to training models on their original loss function and with their original model architecture. This terminology is just convention; many bias mitigation methods are technically also empirical risk minimization. "Importance weighted empirical risk miminization" simply entails modifying one's loss function so that each datapoint (x, y, g) is weighted inversely proportional to the number of datapoints belonging to the same protected group g. This ensures that each protected group has approximately equal representation in the risk function (Saerens et al., 2002).

Adversarial Methods (A) Some of the bias mitigation methods we include are based on robust optimization principles. The *Group Distributionally Robust Optimization (GDRO)* method uses a no-regret algorithm (Hedge) to adversarially sample importance weights w_g for each protected group $g \in \mathcal{G}$ so as to (softly) maximize the current batch loss (Sagawa et al., 2020). This incentives models to prioritize their performance on protected groups on which they currently perform poorly. The *Adversarial Censoring* method also uses an adversary, with the goal of penalizing one's model if its internal representation allows for predicting the protected group of a datapoint (Edwards & Storkey, 2015). This incentives models to learn internal representations that are "blind" to protected group membership.

Unlabeled Methods (U) We also include a number of "unlabeled" bias mitigation methods in our experiments, including Pezeshki et al. (2020)'s *Spectral Decoupling (SD)* and Liu et al. (2021)'s *Just-Train-Twice (JTT)* methods. JTT uses importance weighting to prioritize datapoints that a naively trained model is likely to err on. It first trains an initial model without bias mitigation and then trains a new model, this time upweighting datapoints if the initial model misclassified them. SD uses regularization to incentivize models to not overfit to the use of easy-to-learn features. One shortcoming of SD is that it is intractable to tune its hyperparameters for classification tasks with many classes; even tuning its hyperparameters for binary classification required 800 GPU hours (Pezeshki et al., 2020). As such, we only evaluate the SD method on binary classification tasks on CelebA where we can tune SD's hyperparameters. We also include SD for completeness on our ImageNet experiments using default hyperparameters.

Other Methods Other bias mitigation methods we study also involve modifying a model's architecture. The *Domain Independent* and *Domain Discriminative* methods train multiple copies of a model, with each copy specialized for a specific protected group (Wang et al., 2020). In other words, these methods seek equal representation of each protected group in a model's parameters via segregation. *Invariant risk minimization (IRM)* is a family of methods inspired by causal learning aimed at learning internal representations that are invariant to protected groups, or confounding variables more generally (Arjovsky et al., 2020). In our experiments, we include Arjovsky et al. (2020)'s IRMv1 implementation of IRM.

Our experiments find that, even with significant hyperparameter tuning and custom modifications, the IRM and Adversarial Censoring methods do not perform competitively with other bias mitigation methods and baselines. Thus, we only include these methods in Appendix figures.

4 Revisiting Bias Mitigation Experiments on CelebA

In this section, we analyze the performance of bias mitigation methods on the CelebA dataset (Liu et al., 2015). We extend prior experiment designs by Wang et al. (2020), Liu et al. (2021), and Shrestha et al. (2021) to include a wider range of algorithmic fairness tasks and further analyze bias mitigation methods.

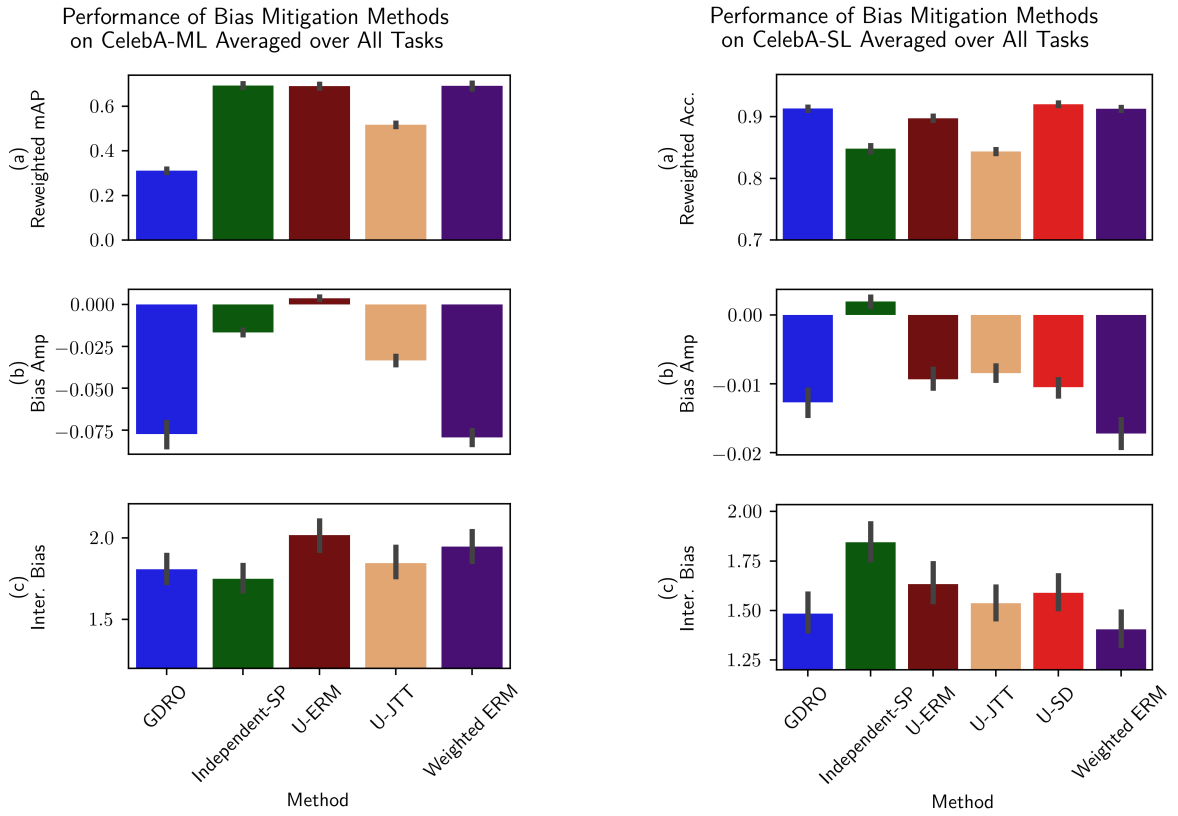
4.1 Experimental Setup

Two common "base" learning tasks on CelebA include the binary classification task *CelebA-SL* where one predicts CelebA's "blond hair" attribute, and the multi-label classification task *CelebA-ML* where one predicts all of CelebA's 40 available attributes. Previous works on bias mitigation have often differed in which base learning task their experiments build on, e.g. Liu et al. (2021) and Shrestha et al. (2021) use CelebA-SL while Wang et al. (2020) uses CelebA-ML. For completeness, we include both in our experiments.

CelebA experiments in prior works design and evaluate bias mitigation methods on a single algorithmic fairness task: choosing a base learning task to solve and a single protected attribute to be fair to. In contrast, our analysis sweeps over 54 different fairness tasks. That is, on both CelebA-SL and CelebA-ML, we run bias mitigation experiments for 27 different choices of $\mathcal{G}$. Recall that $\mathcal{G}$ specifies the set of "protected groups" we want our learned model h to be fair with respect to, so a different choice of $\mathcal{G}$ entail a different fairness task. In each experiment, we apply the bias mitigation methods listed in Table 1 to the training of ResNet-50s (He et al., 2016) and evaluate them on the metrics from Section 3.1. We defer precise hyperparameter and training procedure details to the Appendix, Section B.3.

27 choices of $\mathcal{G}$s We generate our 27 choices of $\mathcal{G}$ using four collections of protected attributes: *Balanced*, *Imbalanced*, *Inconsistent*, and *Protected*. The *Balanced* and *Imbalanced* collections consist of 7 most balanced protected and the 7 least balanced CelebA attributes, respectively. The *Inconsistent* collection consists of 6 inconsistently labeled attributes noted by Ramaswamy et al. (2021). The *Protected* collection consists of 7 legally or societally sensitive attributes. Given an attribute collections $\{a_i\}_{i=1}^k$, for example the *Balanced* collection with $k = 7$ attributes ordered $a_1, \ldots, a_7$, we generate k choices of $\mathcal{G}$s. Specifically, for each $i \in [k]$, we include the set of protected groups $\mathcal{G}^{(i)}$ that arise from the intersection of the first i attributes in our collection: $\mathcal{G}^{(i)} \cong \mathcal{A}_{a_1} \times \cdots \times \mathcal{A}_{a_i}$. This way, the fairness tasks we generate $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(k)}$ not only capture the behavior of bias mitigation methods on different choices of protected attributes, but also on different degrees of intersectionality.

Figures We aggregate the results from all CelebA experiments in Figure 2. These bar plots summarize the experiment results for our 27 choices of $\mathcal{G}$, depicting the mean and standard deviation of three metrics (y-axes): importance-weighted accuracy/mAP, bias amplification scores (Zhao et al., 2017), and intersectional bias scores Foulds et al. (2020). In Figures 3, we instead plot performance metrics for each bias mitigation method against specific groups of experiments. In Figures 3(i)(d,e,f) and Figures 3(ii)(d,e,f), we group experiment settings in terms of intersectionality: if $\mathcal{G}$ arises from the intersections of 3 protected attributes, we plot its experiment results at $x = 3$. This allows us to, for instance, identify which bias mitigation methods are most effective for intersectional fairness settings. In Figures 3(i)(a,b,c) and Figures 3(ii)(a,b,c), we group experiment settings by the performance of learning without bias mitigation. If, for a choice of $\mathcal{G}$, models learned without bias mitigation have an average bias amplification score of 0.01, we plot experiment results for this $\mathcal{G}$ at $x = 0.01$. This allows us to, for instance, identify which bias mitigation methods are most effective in settings where unmitigated learning results in exceptionally great bias. In the Appendix, we provide additional CelebA experiments, including replications of experiments in prior works, and figures, including analogies of Figure 3 for specific attribute collections (e.g. *Inconsistent*).



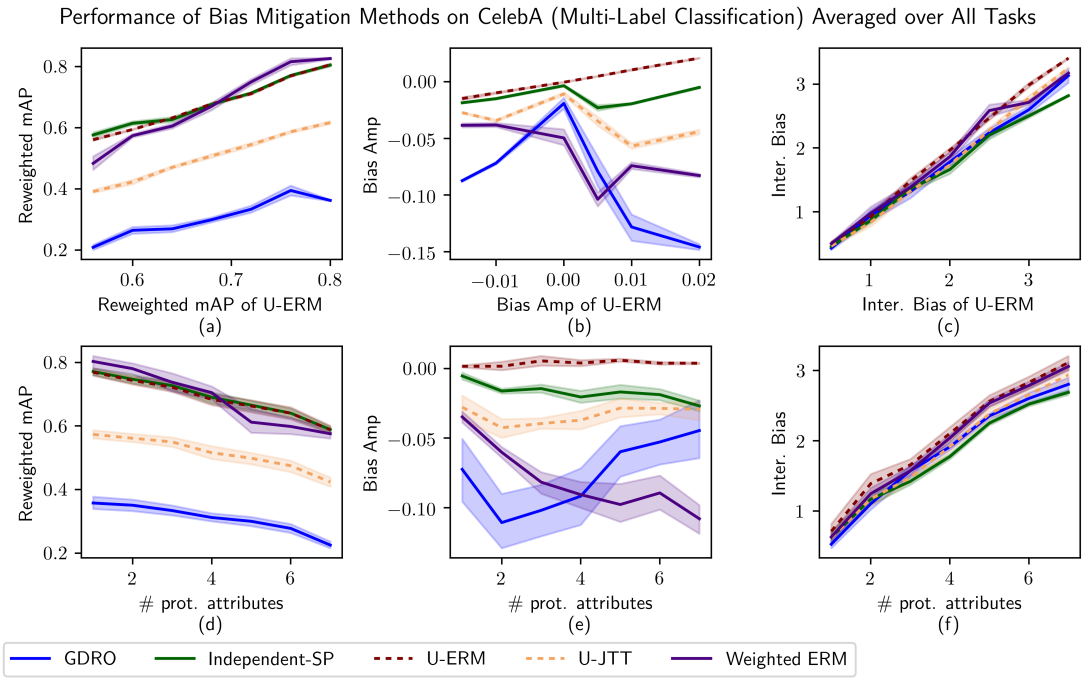
(i) Bias mitigation results for the multi-label classification of celebrity attributes on CelebA.

(ii) Bias mitigation results for the binary classification of hair color on CelebA.

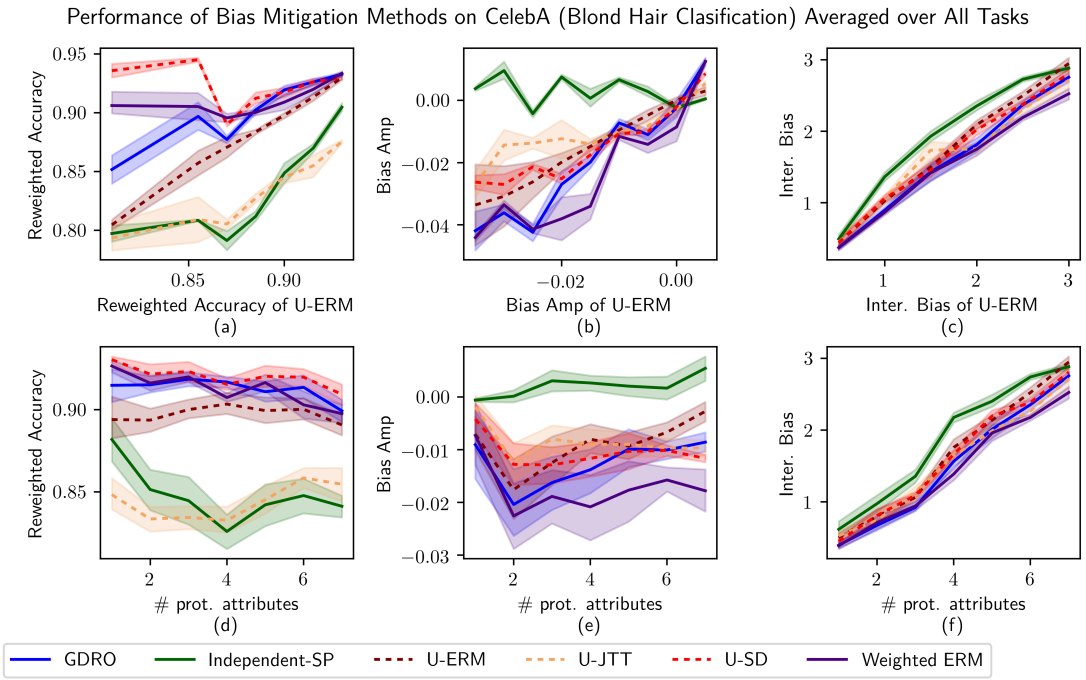
Figure 2: The average performance of ResNet50 models trained using bias mitigation methods on the CelebA dataset. Twenty seven different experiment settings, each designating different sets of protected groups, are represented. Error bars denote 68% confidence intervals. All results are on a test split. Greater reweighted mAP (mean average precision) and reweighted acc (accuracy) indicate more precise/accurate predictions. Greater bias amp (bias amplification) and inter bias (intersectional bias) imply that models score poorly on quantitative estimates of algorithmic bias.

4.2 Key Findings

The outcomes of a CelebA bias mitigation experiment varies greatly depending on one's choice of protected groups $\mathcal{G}$. This phenomenon is particularly noticeable by comparing the plots of bias amplification and intersectional bias scores in Figures 2(i)/2(ii) with those of Figures 3(i)/3(ii). In Figures 2(i)/2(ii), large confidence intervals make it hard to distinguish between the effectiveness of different methods in a statistically significant fashion. This is because Figure 2 averages results over all choices of $\mathcal{G}$, which confounds the performance of each method with the variable difficulty of each experiment setting. In contrast, when we unpack our choices of $\mathcal{G}$ and group them in terms of difficulty and intersectionality in Figures 3(i)/3(ii), we do observe statistically significant trends. We can further observe that the *relative* differences in effectiveness between bias mitigation methods vary depending on which fairness tasks one studies. For instance, in Figure 3(i)(e), we see that the Group DRO method results in lower bias amplification scores on average than simple importance weighted ERM (W-ERM) for non-intersectional choices of $\mathcal{G}$, but this trend reverses for intersectional choices of $\mathcal{G}$. Our findings suggest that the evaluation of bias mitigation methods on a singular CelebA fairness task do not give a complete picture and—as we describe in the Appendix—may misleadingly suggest that certain bias mitigation methods outperform others when such is not the case.

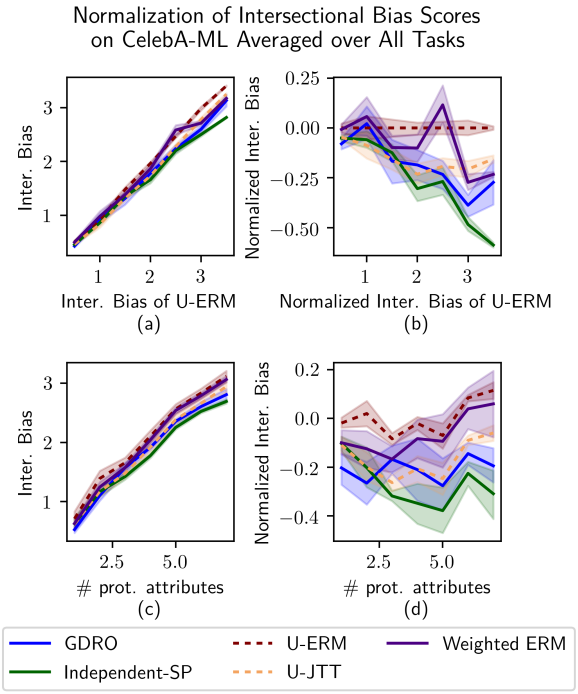


(i) Bias mitigation results for the multi-label classification of celebrity attributes on CelebA.

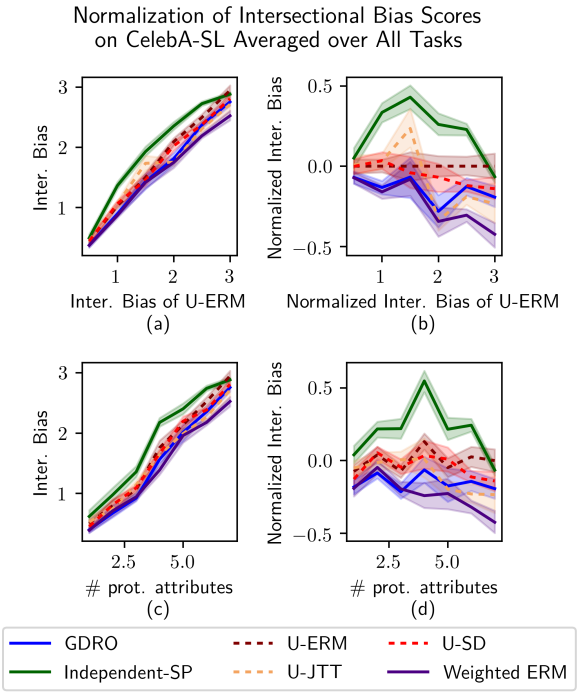


(ii) Bias mitigation results for the binary classification of hair color on CelebA.

Figure 3: The average performance of ResNet50 models trained using bias mitigation methods on the CelebA dataset. Error bars denote 68% confidence intervals. All results are on a test split. These figures depict the same learning tasks, experiment settings, and metrics as Figure 2. In contrast to Figure 2, these figures spread metrics from different experiment settings along the x-axis. The top rows (a-c) plot the performance metrics of different methods against the metrics of learning without bias mitigation (U-ERM) in the same experiment setting. The bottom rows (d-f) plot the performance metrics of different methods against how intersectional an experiment setting is, i.e., the number of protected attributes.



(i) Replottings of Figures 3(i)(c,f) with normalization.



(ii) Replottings of Figures 3(ii)(c,f) with normalization.

Figure 4: The average intersectional bias scores of ResNet50 models trained using bias mitigation methods on the CelebA dataset. Error bars denote 68% confidence intervals. All results are on a test split. These figures replot the intersectional bias figures from Figures 3(i)(c,f) and 3(ii)(c,f) using normalized intersectional bias scores to better visualize the difference between bias mitigation methods. We normalize these plots by, at each increment of the x-axis, offsetting all y-values by the average intersectional bias of learning without mitigation (U-ERM).

Importance weighted ERM remains a consistently effective method of bias mitigation. Whereas the effectiveness of other bias mitigation methods vary significantly based on which CelebA fairness task one looks at, importance weighted ERM (W-ERM) consistently outperforms or ties all other methods in every metric and for almost all choices of $\mathcal{G}$ (Figures 3(i) and 3(ii)). This is surprising as simple importance weighting is one of the earliest approaches to bias mitigation (Saerens et al., 2002), and again highlights the importance of evaluating methods on multiple choices of $\mathcal{G}$.

The Group DRO method performs almost as well as importance weighted ERM on CelebA-SL tasks in all three metrics (Figures 3(ii)(a-e)); however, the method is less effective at bias mitigation in intersectional settings (Figures 3(ii)(e,f)). The Domain Independent method performs at least as well as importance weighted ERM on CelebA-ML tasks in terms of reweighted mAP and intersectional bias scores (Figures 3(i)(a,c,d,f)); however, the method is less effective at reducing bias amplification (Figures 3(i)(b,e)). On the other hand, Group DRO and Domain Independent methods are less effective on CelebA-ML (Figures 3(i)(a,d)) and CelebA-SL (Figures 3(ii)(a,d)) tasks respectively.

The Spectral Decoupling method is surprisingly effective for a method that requires no protected attribute labels. It achieves the highest reweighted accuracy/mAP on CelebA-SL tasks (Figure 3(ii)(a,d)), although it does not significantly reduce bias amplification and intersectional bias scores (Figure 3(ii)(b,c,e,f)). In addition, the approach is of limited practicality beyond binary classification, as its number of hyperparameters scales linearly in the size of one's label space ($|\mathcal{Y}|$) and even tuning the method for just binary classification (where $|\mathcal{Y}| = 2$) requires 800 GPU hours. The Just Train Twice method is uncompetitive across the board (Figures 3(i)(a-e) and 3(ii)(a-e)).

Bias amplification scores and importance weighted accuracy/mAP are non-monotone. One may expect that the performance of a bias mitigation method on a set of fairness tasks should be generally monotone in the difficulty or intersectionality of said tasks. In other words, one would expect that the plots in Figure 3 be monotone. This is indeed the case for intersectional bias scores (Figures 3(i)(c,f), 3(ii)(c, f) and reweighted mAP (Figures 3(i)(a,c)). However, many others, e.g. Figures 3(ii)(d,e), have no discernible pattern along the x-axis. We can rule out the role of noise in such monotonicity as we can observe kinks that exceed our confidence interval. Instead, we deduce that the x-axis (i.e., the difficulty or intersectionality of our experiment settings) does not fully account for variations in the performance of our methods. We qualitatively observe that there is more regularity in Figures 3(i)(a,b), 3(ii)(a,b) than Figures 3(i)(d,e), 3(ii)(d,e), suggesting that variations between different experiment settings are better explained by variations in the performance of unmitigated learning than by intersectionality. This finding is intuitive, as the challenges posed by intersectionality should already be partially captured by the performance of models learned without bias mitigation.

5 Bias Mitigation Methods on ImageNet

In this section, we extend our critical analyses of bias mitigation methods to the ImageNet dataset, in particular, the People Subtree (Deng et al., 2009; Yang et al., 2020). This dataset is significantly larger and more challenging than CelebA, with less than 10% of images have been annotated with protected attributes.

Experimental Setup. The ImageNet People dataset consists of 124,693 images of humans, each labeled with a "synset", an ImageNet term for image category. Examples of synsets on the People subtree include "programmer" and "child". Recently, Yang et al. (2020) annotated datapoints from the ImageNet People subtree with 3 protected attributes: gender, skin color, and age. With these annotations, Yang et al. (2020) identified numerous instances of the under-representation of protected groups in ImageNet. The intersections of these attributes compose 196 intersectional groups, i.e., $|\mathcal{G}| = 196$, which we will use to design a challenging real-world benchmark on which to evaluate bias mitigation methods. An important aspect of this benchmark is that only 15,981 images in the dataset have been labeled with protected attributes; this poses a serious challenge with attribute label scarcity.

We divide our dataset into a training split with 124,693 images (of which 5,861 have protected attribute labels), and validation and test splits with 5,327 images each (all of which have protected attribute labels). In our experiment, we apply the bias mitigation methods listed in Table 1 to the training of ResNet50s (He et al., 2016) on our ImageNet training set. We then evaluate them on metrics from Section 3.1.

Most of the bias mitigation methods we study depend on access to protected group labels, rendering most of our ImageNet training set unusable for them. As such, in our experiment we first pretrain each of our ResNet50 models using empirical risk minimization with standard risk functions (i.e., no bias mitigation). We then use our bias mitigation methods to finetune the pretrained models. In Appendix A.3, we show that omitting this pretraining stage results in poor performance.

Findings Table 2 summarizes our evaluations of bias mitigation methods on the ImageNet dataset. Overall, we observe similar qualitative trends as Figure 3(i) in terms of which methods are most effective. In particular, Importance Weighted ERM and the Domain Independent method are most effective and improve upon un-mitigated learning (U-ERM) in terms of average reweighted accuracy (47%, 47% vs 45%) and bias amplification (7.6, 7.4 vs 8.2).

However, we highlight that the improvements afforded by the use of these bias mitigation methods are not significant on the ImageNet dataset. Even the reductions in bias amplification afforded by Importance Weighted ERM and the Domain Independent method fall within a standard deviation in Table 2. This is in contrast to CelebA experiments, where these same bias mitigation methods resulted in statistically significant reductions in bias metrics on a diverse range of fairness tasks. In the following sections, we show that this is due to label-scarcity and identify several techniques for improving the effectiveness of bias mitigation methods in ImageNet's label-scarce setting.

Table 2: The average performance of ResNet50 models trained using bias mitigation methods on the ImageNet dataset. Standard deviation is indicated by ±. Importance weighted ERM (W-ERM) and the Domain Independent method (Ind) are the most effective bias mitigation techniques. However, no bias mitigation method is significantly better than naively training without bias mitigation (U-ERM).

	Reweighted Acc.	Bias Amp. 100x	Inter. Bias
U-ERM	45.38 ± 0.404	8.153 ± 0.95	$\underline{2.684} \pm 0.03$
UWA-JTT	45.42 ± 0.460	$\underline{7.603} \pm 1.05$	2.694 ± 0.01
U-SD	36.59 ± 0.352	8.489 ± 1.16	$\mathbf{2.621} \pm 0.03$
WA-GDRO	45.20 ± 0.136	8.910 ± 1.70	2.697 ± 0.01
W-ERM	$\mathbf{47.67} \pm 0.812$	7.611 ± 0.11	2.687 ± 0.03
Ind	$\underline{47.24} \pm 0.817$	$\mathbf{7.402} \pm 0.95$	2.688 ± 0.05

5.1 Scaling by Reducing Model Complexity

We identify two techniques for improving the effectiveness of bias mitigation methods in ImageNet's label-scarce setting: *knowledge distillation* and *attribute decomposition*. We apply them to two methods that we identified as particularly effective in Table 2: Importance Weighted ERM and the Domain Independent method. Intuitively, *knowledge distillation* and *attribute decomposition* follow the principle of parsimony and mitigate the effects of label scarcity by reducing the internal complexities (e.g., degrees of freedom) of bias mitigation methods.

Attribute Decomposition (AD). The complexity of most bias mitigation methods scale linearly with the number of protected groups, and thus exponentially in the number of protected attributes. *Attribute Decomposition* modifies bias mitigation methods so that their degrees of freedom grow linearly with the number of protected attributes. One drawback of AD is that it limits the ability of bias mitigation methods to address intersectional effects. We now describe AD implementations of Importance Weighted ERM and the Domain Independent method on a hypothetical learning problem defined by the tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{G}, \mathcal{A}, \text{Attrs})$, as per Section 3.

The Domain Independent (Ind) method (Wang et al., 2020) trains a specialized model h_g for each protected group $g \in \mathcal{G}$. Each of these models learns the same original $\mathcal{X} \to \mathcal{Y}$ task, but only trains on datapoints corresponding to their group. During inference, given an input x, the method sums the outputs of every specialized models, outputting $\sum_{g \in \mathcal{G}} h_g(x)$. The AD variant of the Domain Independent method trains, for every protected attribute $a \in \text{Attrs}$ and for every value $v \in \mathcal{A}_a$ that a may take, a specialized model $h_{a,v}$. This model $h_{a,v}$ only trains on datapoints whose protected attribute a takes the value v. During inference, given an input x, our AD method sums the outputs of these models: $\sum_{a \in \text{Attrs}} \sum_{v \in \mathcal{A}_a} h_{a,v}(x)$.

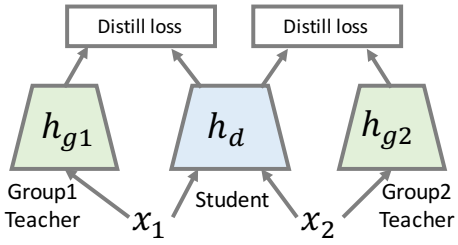
Importance Weighted ERM modifies one's loss function so that each datapoint (x, y, g) is weighted by a group-specific weight $w_g := 1/\Pr_{g' \sim S_\mathcal{G}}(g' = g)$, where $S_\mathcal{G}$ is the empirical distribution of protected groups over the training dataset. The AD variant of Importance Weighted ERM specifies, for every protected attribute $a \in \text{Attrs}$, an importance weight for every value $v \in \mathcal{A}_a$ that a may take: $w_{a,v} = 1/\Pr_{v' \sim S_a}(v' = v)$. Here, S_a is the empirical distribution of values that the protected attribute a takes over the training dataset. A datapoint with protected attribute values $v \in \mathcal{A}$ is then given the importance weight $\prod_{a \in attrs} w_{a,v_a}$ in the loss function.

Knowledge Distillation (KD). Some bias mitigation methods rely on training different models for each protected group (Wang et al., 2020). For instance, recall that the Domain Independent method trains multiple copies of a model. *Knowledge Distillation* modifies these methods by consolidating these multiple copies back into a single model. Specifically, we can treat these model copies $\{h_g \mid g \in \mathcal{G}\}$ as teacher classifiers and apply knowledge distillation (Hinton et al., 2015) to learn a single student classifier h_d using the following loss function:

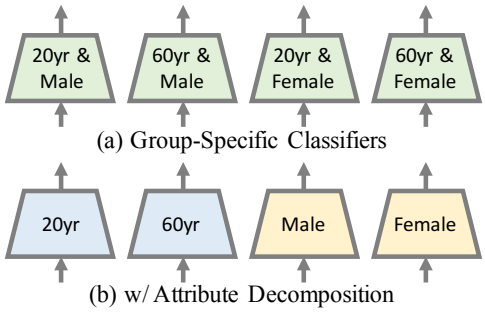
$$\ell(h_d(x), y) + \lambda KL(h_d(x), h_g(x)). \tag{3}$$

Here, h_g are already trained models, $\ell(\cdot, \cdot)$ is our loss function and λ is the weight for the distillation term.

Findings. Table 3 compares Importance Weighted ERM and the Domain Independent method with their Attribute Decomposition and Knowledge Distillation counterparts. The fourth row describes an



Knowledge Distillation (KD): Some bias mitigation methods train multiple classifiers, depicted by h_{g_1} and h_{g_2}. KD consolidates these classifiers into a single "student" classifier, depicted by h_d. The *distillation loss* trains the student h_d to, given datapoints x_1, x_2, predict the same labels as those predicted by h_{g_1} and h_{g_2}.



Attribute Decomposition (AD): Some bias mitigation methods train specialized classifiers for each protected group (a). AD instead trains specialized classifiers for each protected attribute value (b).

Table 3: The average performance of ResNet50 models trained using bias mitigation methods on the ImageNet dataset. Standard deviation is indicated by ±. Knowledge Distillation (KD) and Attribute Decomposition (AD) each improve the effectivess of bias mitigation methods by a statistically significant margin.

	Reweighted Acc.	Bias Amp. 100x	Inter. Bias
Ind	47.24 ± 0.817	7.402 ± 0.95	2.688 ± 0.05
w/ AD	45.12 ± 0.107	6.906 ± 0.24	$\mathbf{2.621} \pm 0.03$
w/ KD	$\mathbf{47.68} \pm 0.521$	$\underline{6.318} \pm 1.00$	2.693 ± 0.01
w/ KD & AD	47.47 ± 0.323	$\mathbf{5.901} \pm 0.73$	2.689 ± 0.01
W-ERM	$\underline{47.67} \pm 0.812$	7.611 ± 0.11	$\underline{2.687} \pm 0.03$
w/ AD	47.61 ± 0.478	6.843 ± 0.64	2.709 ± 0.03

implementation of the Domain Independent method using both AD and KD. AD and KD significantly improve the effectiveness of the two bias mitigation methods, particularly as measured by Bias Amplification scores. KD reduces bias amplification for Domain Independent ($7.4 \rightarrow 6.3$) while maintaining accuracy metrics. AD reduces bias amplification for both methods ($7.4 \rightarrow 6.9$ for Ind and $7.6 \rightarrow 6.8$ for W-ERM). In addition, combining AD+KD yields the best bias amplification score of 5.9. While we previously found in Table 2 that no bias mitigation method reduced bias amplification scores by a statistically significant quantity, the use of Attribute Decomposition and Knowledge Distillation leads to a statistically significant reduction in bias amplification from 8.2 ± 1 to 5.9 ± 0.7.

6 Conclusion

We have proposed a broader analysis of bias mitigation methods on the commonly used CelebA dataset and the use of the ImageNet dataset as a test-bed for label scarce bias mitigation. Our analyses reveal that the high reported performance of many bias mitigation algorithms is due to somewhat arbitrary choices in experiment design, and in fact plain importance weighting remains empirically the most effective and reliable bias mitigation method. We have proposed two techniques for adapting bias mitigation algorithms to the label-scarce ImageNet dataset for evaluation. We hope this work encourages a more critical review of popular bias mitigation methods and the future use of more diverse sets of empirical benchmarks.

We now highlight several future research directions:

- Identify potential mechanisms for the incredible robustness of simple importance weighting.
- Gain a better understanding for why otherwise effective bias mitigation methods break down when evaluated on a more diverse set of tasks.
- Explore other schemes for overcoming sparsity in protected attribute labels.

References

Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. Fairness and Robustness in Invariant Learning: A Case Study in Toxicity Classification. *arXiv:2011.06485 [cs]*, December 2020. URL `http://arxiv.org/abs/2011.06485`. arXiv: 2011.06485.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, March 2020. URL `http://arxiv.org/abs/1907.02893`.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 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.

Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms, July 2020. URL `http://arxiv.org/abs/2007.06570`. arXiv:2007.06570 [cs].

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

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

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 339–348, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287594. URL `https://doi.org/10.1145/3287560.3287594`.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, pp. 139, 1989. Publisher: HeinOnline.

Enyan Dai and Suhang Wang. Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pp. 680–688, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8297-7. doi: 10.1145/3437963.3441752. URL `https://doi.org/10.1145/3437963.3441752`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image Counterfactual Sensitivity Analysis for Detecting Unintended Bias, October 2020. URL `http://arxiv.org/abs/1906.06439`. arXiv:1906.06439 [cs, stat].

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. In *ICLR*, November 2015. URL https://arxiv.org/abs/1511.05897v3.

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*, pp. 259–268, 2015.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921. IEEE, 2020.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, April 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/content/115/16/E3635.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

Caroline Ho, Hugo Kitano, and Kevin Lee. Fair Image Classification with Semi-Supervised Learning. Technical report, Stanford, 2020.

Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, April 2021. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100241. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8085589/.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 110, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3373154. URL https://doi.org/10.1145/3351095.3373154.

Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. MultiFair: Multi-Group Fairness in Machine Learning. *arXiv:2105.11069 [cs, math, stat]*, May 2021. URL http://arxiv.org/abs/2105.11069.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572. PMLR, 2018.

Youjin Kong. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 485–494, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533114. URL https://dl.acm.org/doi/10.1145/3531146.3533114.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 2017.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. *ICML*, 2021. URL http://arxiv.org/abs/2107.09044.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 349–358, 2019.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally-motivated Shortcut Removal Using Auxiliary Labels. *arXiv:2105.06422 [cs]*, June 2021. URL http://arxiv.org/abs/2105.06422.

Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. Algorithms of oppression: How search engines reinforce racism. New York University Press, New York, NY, US, 2018. ISBN 978-1-4798-3724-3 978-1-4798-4994-9.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. *arXiv:2011.09468 [cs, math, stat]*, November 2020. URL http://arxiv.org/abs/2011.09468.

Vikram V. Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9301–9310, 2021.

Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity. *arXiv:1712.00193 [cs]*, July 2018. URL http://arxiv.org/abs/1712.00193.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, January 2002. ISSN 0899-7667. doi: 10.1162/089976602753284446.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *ICLR*, 2020. URL http://arxiv.org/abs/1911.08731.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. An Investigation of Critical Issues in Bias Mitigation Techniques. *arXiv:2104.00170 [cs, stat]*, March 2021. URL http://arxiv.org/abs/2104.00170.

Latanya Sweeney. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29, 2013. Publisher: ACM New York, NY, USA.

Angelina Wang and Olga Russakovsky. Directional bias amplification. *ICML*, 2021.

Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. *arXiv preprint arXiv:2205.04610*, 2022.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

A Appendix: Additional Experiments and Figures

A.1 Replication Experiments

In this section, we replicate experiment designs considered in previous works and extend their results to include additional methods and metrics.

Table 4: Shrestha et al. (2021) Replication (CelebA-SL)

Model	Inter. Bias	Bias Amp	Accuracy	Reweighted Accuracy	Min Accuracy
U-ERM	1.005 ± 0.109	-0.024 ± 0.012	0.923 ± 0.012	0.848 ± 0.012	0.604 ± 0.065
WA-GDRO	0.558 ± 0.049	-0.065 ± 0.008	0.911 ± 0.010	0.895 ± 0.009	0.833 ± 0.034
IRMv1	1.071 ± 0.148	-0.010 ± 0.010	0.940 ± 0.007	0.793 ± 0.009	0.404 ± 0.013
U-SD	0.746 ± 0.032	-0.057 ± 0.004	0.884 ± 0.005	0.887 ± 0.005	0.817 ± 0.025
W-ERM	0.610 ± 0.069	-0.059 ± 0.008	0.915 ± 0.006	0.903 ± 0.007	0.852 ± 0.041

In Table 4, we replicate the CelebA experiments from Shrestha et al. (2021). This entails training a ResNet18 on CelebA for the task of predicting the attribute "has blond hair" while protecting the attribute "is male". Our figures fall within the margin of error of those reported in Shrestha et al. (2021). Our extended results reveal that while Spectral Decoupling (U-SD) may be effective in improving reweighted and worst-case accuracy, Importance Weighted ERM (W-ERM) and Group Distributionally Robust Optimization (WA-GDRO) are significantly more effective in reducing intersectional bias. These results mostly align with the general trends we observe in Figure 3.

Table 5: Liu et al. (2021) Replication (CelebA-SL)

Model	Inter. Bias	Bias Amp	Accuracy	Reweighted Accuracy	Min Accuracy
U-ERM	1.107 ± 0.606	0.024 ± 0.018	0.930 ± 0.044	0.715 ± 0.148	0.281 ± 0.197
WA-GDRO	0.666 ± 0.010	-0.049 ± 0.002	0.936 ± 0.002	0.916 ± 0.001	0.857 ± 0.003
UWA-JTT	0.911 ± 0.005	-0.041 ± 0.002	0.902 ± 0.005	0.907 ± 0.001	0.842 ± 0.003
W-ERM	0.647 ± 0.101	-0.055 ± 0.012	0.922 ± 0.010	0.921 ± 0.003	0.891 ± 0.026

In Table 5, we replicate the CelebA experiments from Liu et al. (2021). This entails training a ResNet50 on CelebA for the task of predicting the attribute "has blond hair" while protecting the attribute "is male". Our figures fall within the margin of error of those reported in Liu et al. (2021) and confirm their findings that Just Train Twice (UWA-JTT) matches the Group DRO method and Importance Weighted ERM in terms of reweighted and worst-case accuracy. However, our extended results reveal that JTT does not lead to the same improvements in intersectional bias and bias amplification scores as Group DRO and Importance Weighted ERM. Moreover, as seen in Figure 3, the impressive performance of the Just Train Twice method in this experiment setting appears to be an outlier. As seen in Figure 7, even switching the protected and target labels in this experiment setting results in the Just Train Twice method performing no better than the baseline (U-ERM).

Table 6: Wang et al. (2020) Replication (CelebA-ML)

Model	mAP ↑	Reweighted mAP ↑	Inter. Bias ↓	Bias Amp ↓
U-ERM	0.794 ± 0.001	0.746 ± 0.001	1.179 ± 0.037	0.007 ± 0.002
Disc	0.792 ± 0.001	0.739 ± 0.002	1.176 ± 0.012	0.007 ± 0.003
WA-GDRO	0.639 ± 0.003	0.569 ± 0.002	1.316 ± 0.018	0.039 ± 0.003
Ind (w/o sum-prob)	0.780 ± 0.001	0.760 ± 0.000	0.854 ± 0.021	-0.029 ± 0.004
Ind (w/ sum-prob)	0.779 ± 0.003	0.757 ± 0.002	0.837 ± 0.026	-0.025 ± 0.005
A-Cens	0.770 ± 0.001	0.706 ± 0.002	1.324 ± 0.026	0.026 ± 0.002
W-ERM	0.767 ± 0.001	0.772 ± 0.004	0.774 ± 0.007	-0.061 ± 0.003

In Table 6, we replicate the CelebA experiments from Wang et al. (2020). This entails training a ResNet50 on CelebA for the task of predicting 34 CelebA attributes while protecting the attribute "is male". Our figures fall within the margin of error of those reported in Wang et al. (2020) and confirm their findings that their proposed Domain Independent method (Ind) significantly outperforms the Domain Discriminative, Group DRO, and Adversarial Censoring methods. Moreover, their method outperforms them in all metrics

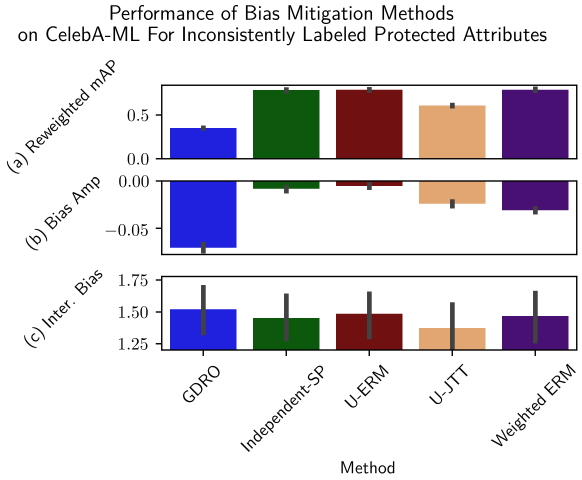
including the intersectional bias score, which we added to their experiment. However, our extended results also introduce Importance Weighted ERM as a baseline and reveal that the Domain Independent method does not improve upon Importance Weighted ERM in any of the bias metrics. We do eventually find in Table 3 that the Domain Independent method is extremely effective when adapted with Knowledge Distillation and Attribute Decomposition and applied to the label-scarce ImageNet dataset.

A.2 Additional Results from Section 4

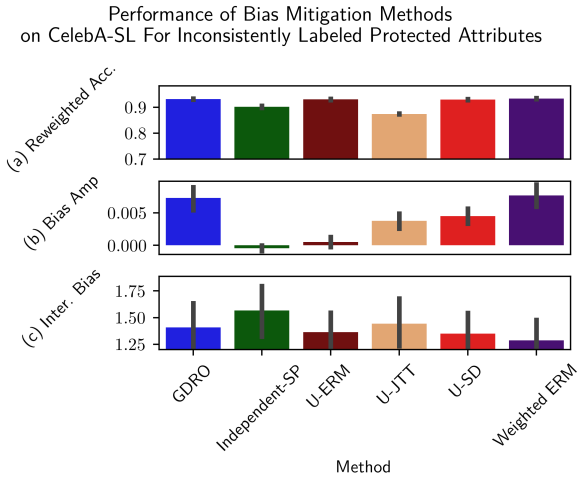
For our CelebA experiments depicted in Figures 2 and 3, we generate 54 different experiment settings by sweeping over four collections of CelebA attributes, as described in Section 3. We describe these four collections in full below:

1. Balanced: Wearing Lipstick, High Cheekbones, Heavy Makeup, Male, Attractive, Smiling, Mouth Slightly Open. These protected attributes are, in order, the CelebA attributes with the most balanced labels (i.e. close to 50-50)
2. Imbalanced: Wearing Hat, Double Chin, Blurry, Gray Hair, Bald, Sideburns, Mustache. These protected attributes are, in order, the CelebA attributes with the most imbalanced labels (i.e. far from 50-50).
3. Inconsistent: Wavy Hair, Oval Face, Big Nose, Pale Skin, Big Lips, Straight Hair. These protected attributes are the CelebA attributes designated as "inconsistently labeled" by Ramaswamy et al. (2021).
4. Protected: Pale Skin, Male, Narrow Eyes, Big Nose, Young, Straight Hair, and Attractive. These attributes are selected for their associations with protected demographics.

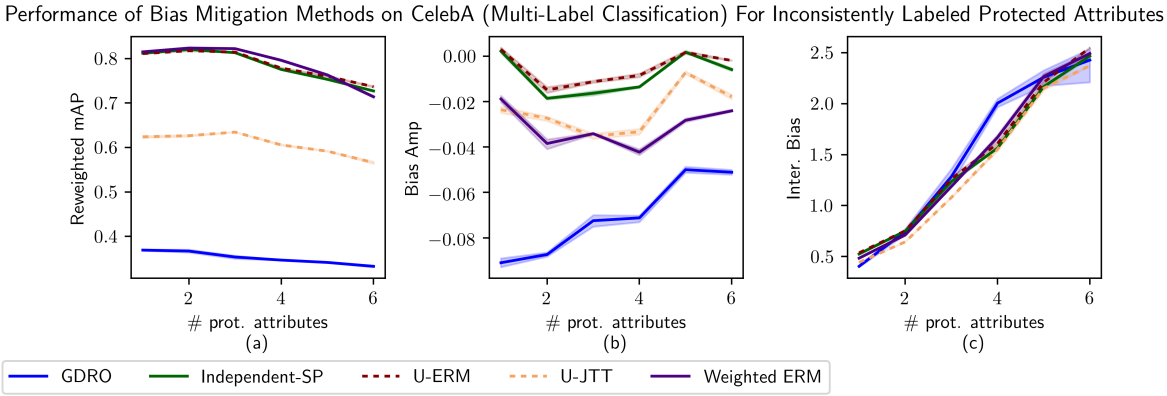
Recall that Figures 2 and 3 aggregate results from the experiment settings generated by all four of these attribute collections. In the following pages, for each attribute collection, we generate analogies of Figures 2 and 3 that only contain experiment settings generated from said collection. The variation between these figures for different choices of attribute collections underscores our earlier discussion about the importance of aggregating diverse experiment settings when evaluating bias mitigation methods. These collection-specific figures may also be of use to readers interested in the behavior of bias mitigation experiments in specific circumstances, e.g. when protected attribute are noisily labeled (see Figure 6).



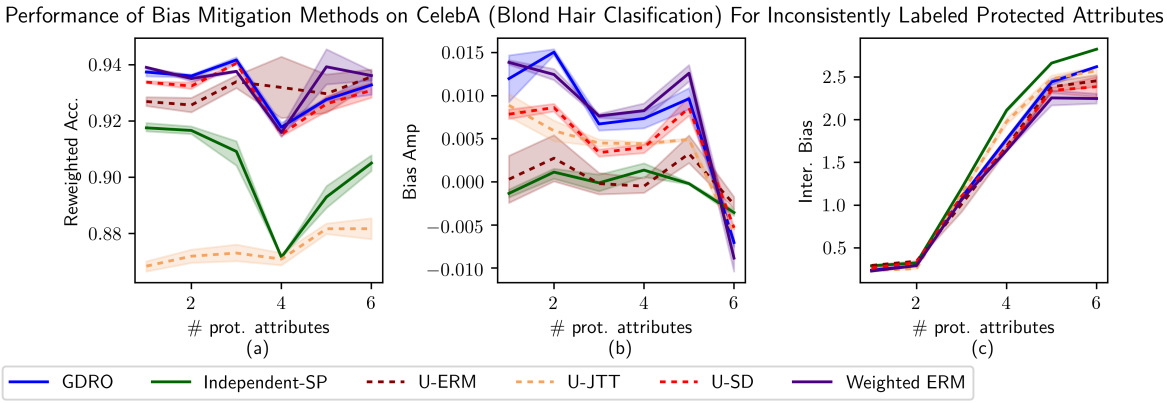
(i) Bias mitigation results for CelebA-ML.



(ii) Bias mitigation results for CelebA-SL.

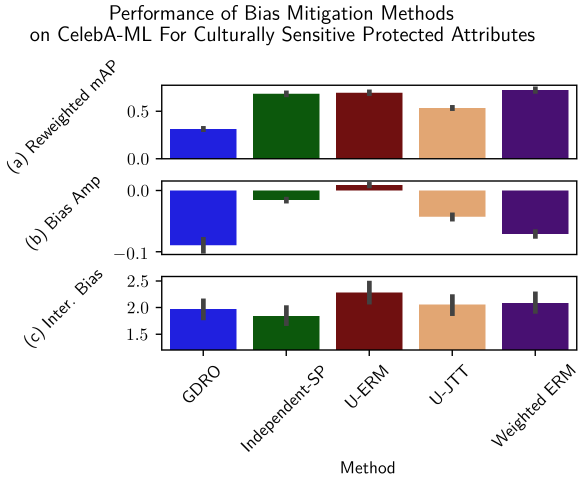


(iii) Bias mitigation results for the multi-label classification of celebrity attributes on CelebA.

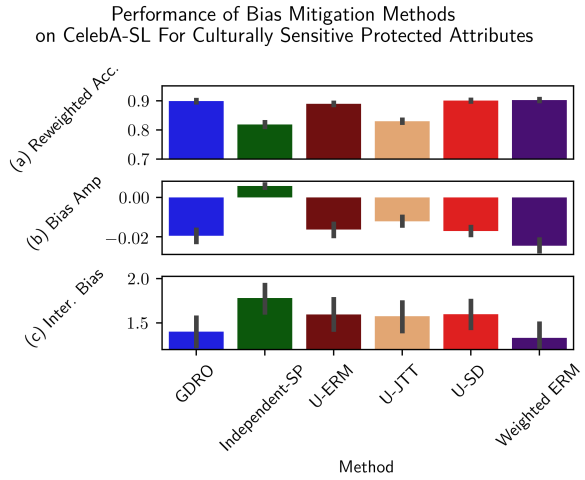


(iv) Bias mitigation results for the binary classification of hair color on CelebA.

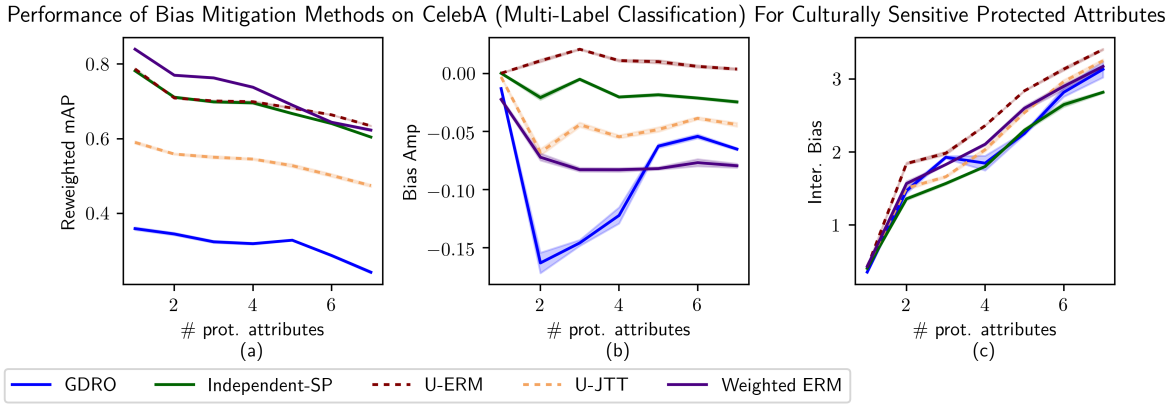
Figure 6: **Six experiment settings generated from the Inconsistent attribute collection:** The average test-split performance of ResNet50 models trained with bias mitigation on CelebA. Error bars denote 68% CI. In (iii-iv), metrics at $x = 1$ correspond to "protecting" the first attribute in the collection. Metrics at $x = 6$ correspond to "protecting" the intersections of every attribute in the collection.



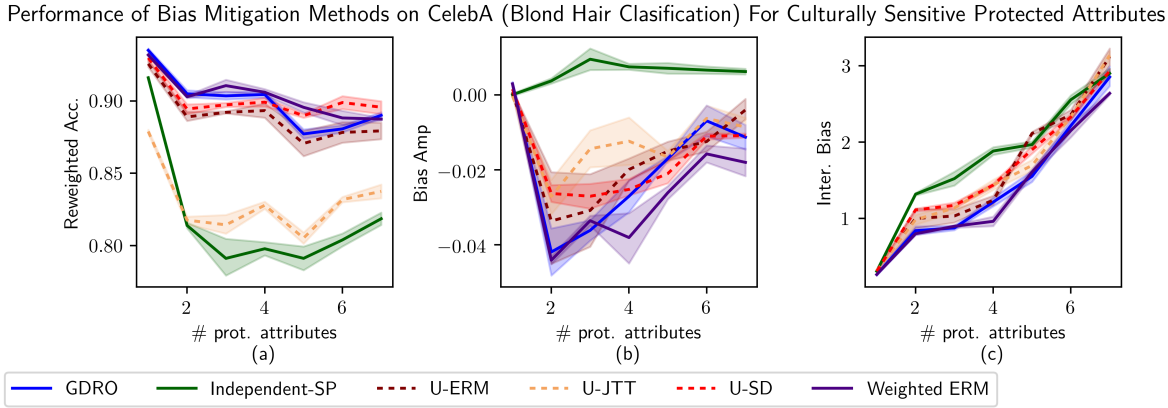
(i) Bias mitigation results for CelebA-ML.



(ii) Bias mitigation results for CelebA-SL.

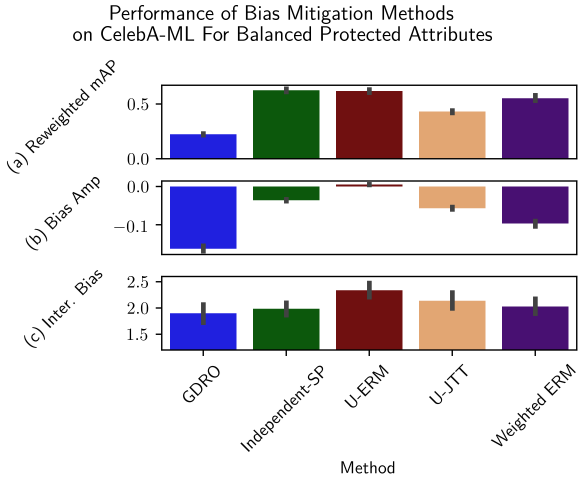


(iii) Bias mitigation results for the multi-label classification of celebrity attributes on CelebA.

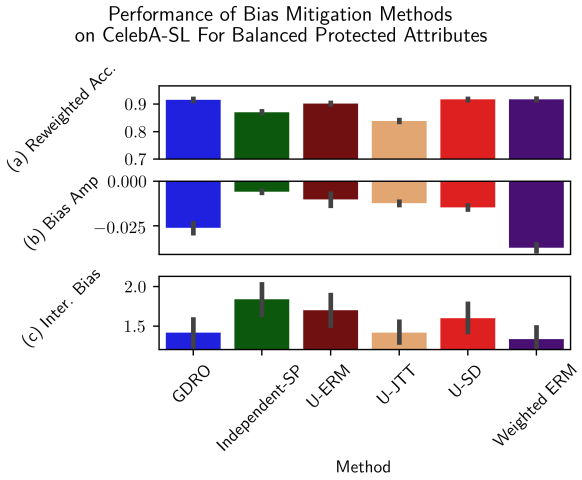


(iv) Bias mitigation results for the binary classification of hair color on CelebA.

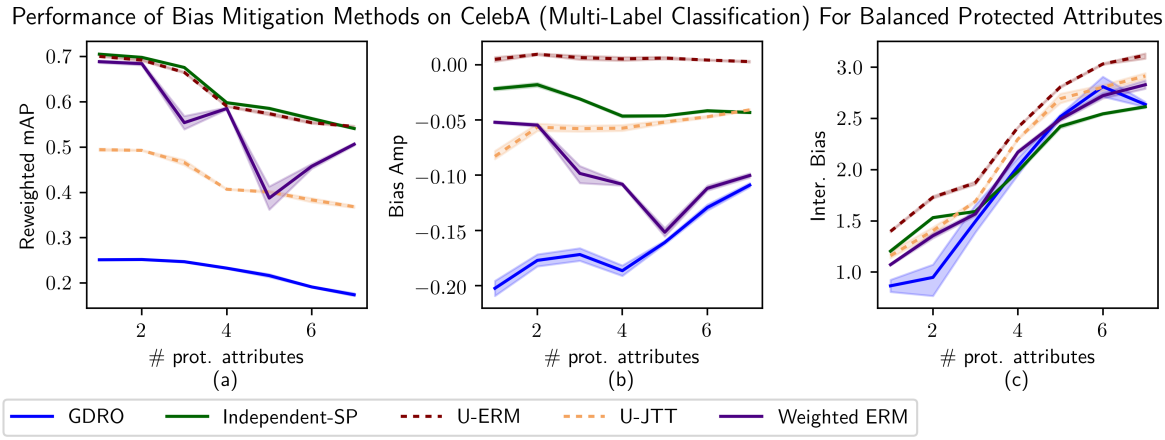
Figure 7: **Seven experiment settings generated from the Protected attribute collection:** The average test-split performance of ResNet50 models trained with bias mitigation on CelebA. Error bars denote 68% CI. In (iii-iv), metrics at $x = 1$ correspond to "protecting" the first attribute in the collection. Metrics at $x = 7$ correspond to "protecting" the intersections of every attribute in the collection.



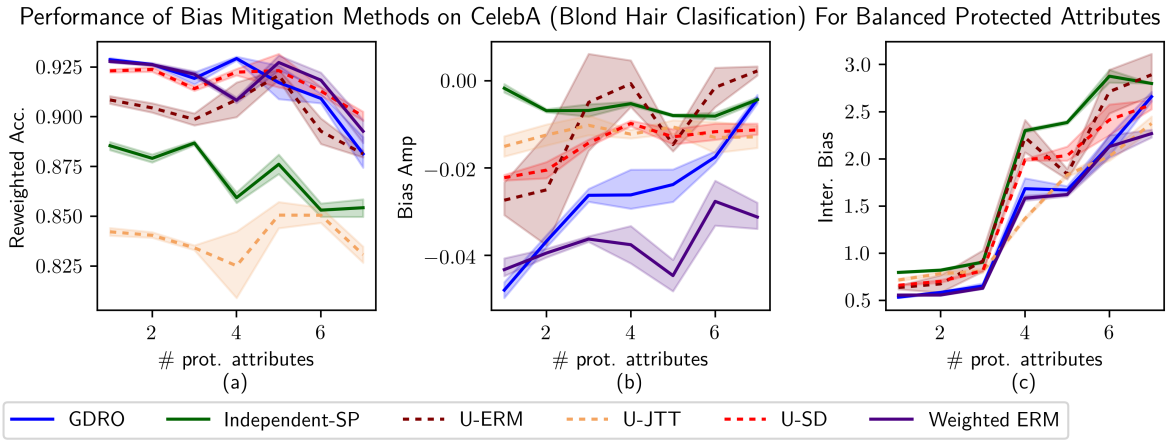
(i) Bias mitigation results for CelebA-ML.



(ii) Bias mitigation results for CelebA-SL.

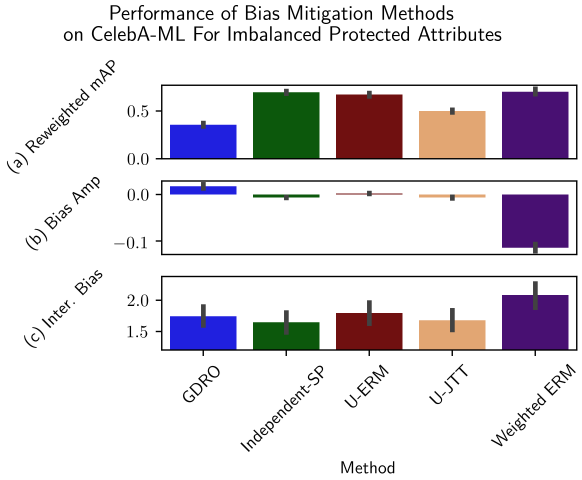


(iii) Bias mitigation results for the multi-label classification of celebrity attributes on CelebA.

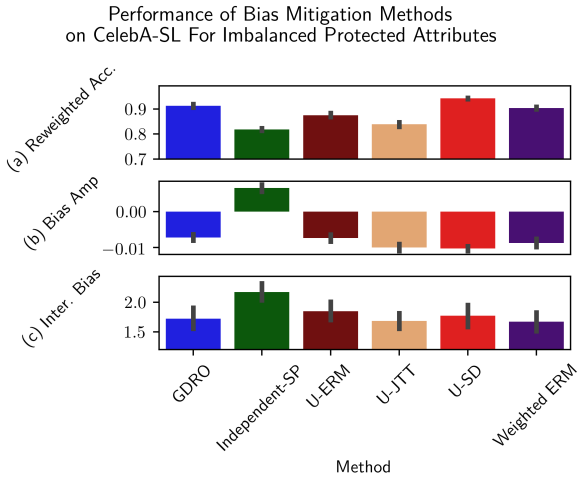


(iv) Bias mitigation results for the binary classification of hair color on CelebA.

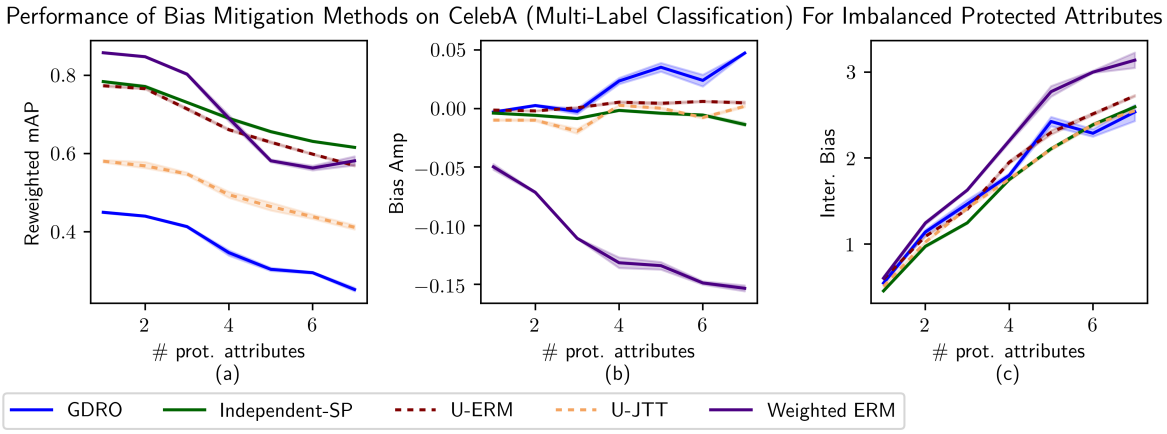
Figure 8: **Seven experiment settings generated from the Balanced attribute collection:** The average test-split performance of ResNet50 models trained with bias mitigation on CelebA. Error bars denote 68% CI. In (iii-iv), metrics at $x = 1$ correspond to "protecting" the first attribute in the collection. Metrics at $x = 7$ correspond to "protecting" the intersections of every attribute in the collection.



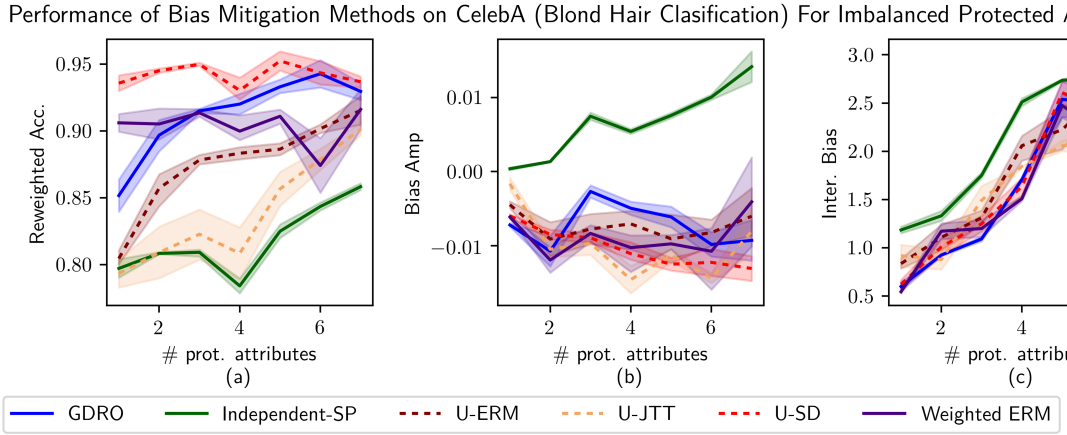
(i) Bias mitigation results for CelebA-ML.



(ii) Bias mitigation results for CelebA-SL.



(iii) Bias mitigation results for the multi-label classification of celebrity attributes on CelebA.



(iv) Bias mitigation results for the binary classification of hair color on CelebA.

Figure 9: **Seven experiment settings generated from the Imbalanced attribute collection:** The average test-split performance of ResNet50 models trained with bias mitigation on CelebA. Error bars denote 68% CI. In (iii-iv), metrics at $x = 1$ correspond to "protecting" the first attribute in the collection. Metrics at $x = 7$ correspond to "protecting" the intersections of every attribute in the collection.

Table 7: Pretraining significantly improves reweighted accuracy and reduces bias amplification. However, pretraining increases intersectional bias as the metric is uninformative for underfitting models.

	Reweighted Acc.	Bias Amp. 100x	Inter. Bias
W-ERM	47.67 ± 0.812	7.611 ± 0.11	2.687 ± 0.03
w/o pretrain	14.88 ± 0.670	9.602 ± 1.54	2.167 ± 0.04
Ind	47.24 ± 0.817	7.402 ± 0.95	2.688 ± 0.05
w/o pretrain	17.36 ± 0.255	14.29 ± 1.10	2.268 ± 0.01

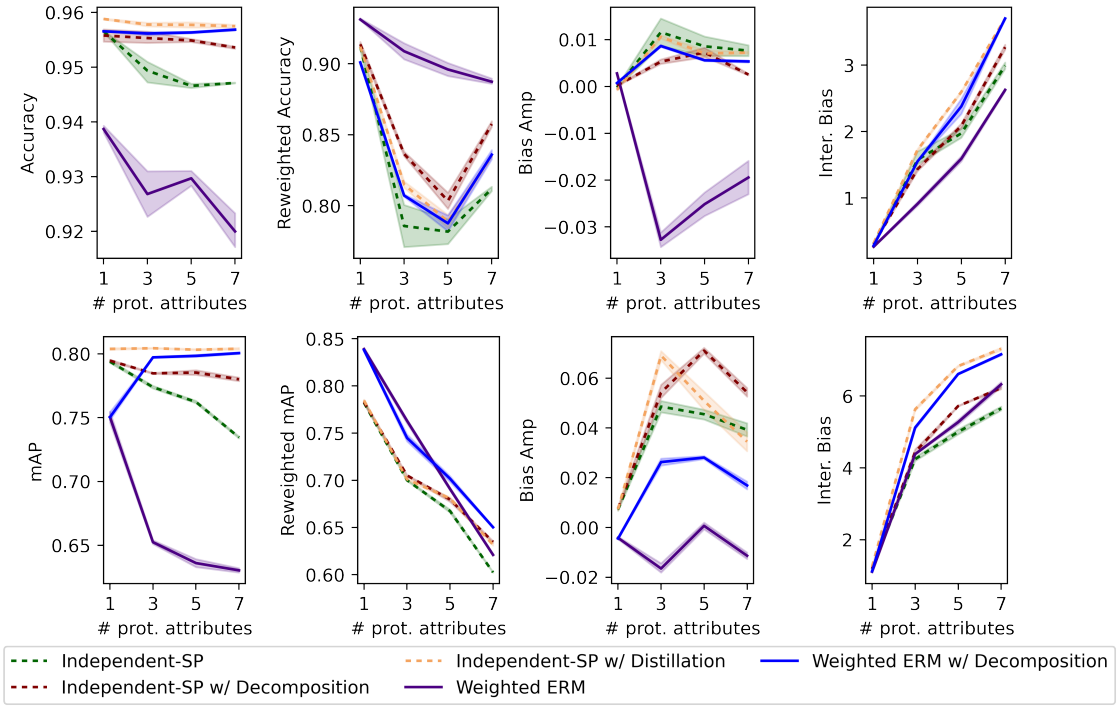


Figure 10: The average performance of ResNet50 models trained using bias mitigation methods on the CelebA dataset for the CelebA-SL task (top) and CelebA-ML task (bottom). These figures represent 7 experiment settings generated from the *Protected* attribute collection. Error bars denote 68% confidence intervals. All results are on a test split.

A.3 ImageNet Pretraining Experiments

In Table 7, we depict the performance of bias mitigation methods on the training of ResNet50 models on the ImageNet dataset. This experiment setting is identical to that plotted Table 2, except that we include the performances of Importance Weighted ERM and the Domain Independent method if they had not used pretraining (denoted as "w/o pretrain"). As expected, pretraining is necessary to achieve meaningful performance using bias mitigation methods on a label-scarce dataset, improving reweighted accuracies from 14% $\rightarrow$ 48% and 17% $\rightarrow$ 47%. As such, we include a pretraining stage as part of the default experiment procedures described in Section 5.

A.4 Knowledge Distillation and Attribute Decomposition CelebA Experiments

In Figure 10, we study the effect of modifying bias mitigation methods with Knowledge Distillation (KD) and Attribute Decomposition (AD) before applying them to the CelebA experiment settings in Figure 7. As expected, we observe that—in contrast to the trends observed in Table 3—KD and AD appear to result in less effective bias mitigation methods. This is because label scarcity is not an issue on CelebA, so reducing the degrees of freedom of bias mitigation methods is counterproductive in this setting. Indeed, we see that the counterproductive effects of KD and AD become statistically significant in intersectional settings (when the # of protected attributes exceeds 3) where oversimplifying bias mitigation methods is most harmful. This

highlights that techniques for bias mitigation that are most effective in label-scarce settings like ImageNet, may not be most effective in label-plentiful settings like CelebA, and vise-versa.

B Appendix: Experiment Details

B.1 Computing Environment

Each experiment run, corresponding to the training of a single model for a single random seed, is trained concurrently with two other runs on their own GPU type Tesla V100-SXM2-32GB-LS provisioned from a commercial cloud service. The training process for three such concurrent runs takes anywhere from two GPU hours up to twenty four GPU hours.

B.2 Dataset Information

Here, we detail the two main datasets that our experiments use. The first is CelebA (Liu et al., 2015), a dataset of celebrity facial pictures. CelebA labels these facial images with forty binary attributes such as "Pale"/"Not Pale" and "Male"/"Not Male"; these binary definitions and choices of default values originate from the dataset itself and not the authors of this work. We have retained the dataset's original terminology for continuity with prior literature. The second is the ImageNet dataset, specifically the People subtree of the ImageNet challenge (Deng et al., 2009). This is a multi-class classification task. We use protected attribute labels provided by Yang et al. (2020) about the gender, skin color and age of each image's photographed individual. These attribute labels only cover around $15,000$ of the $140,000$ images we were able to download from the ImageNet People subtree. Note that this People Subtree is not the usual ILSVRC subset of ImageNet adopted by other computer vision literature. Accuracies on this dataset, e.g. in Table 2, may appear low but are only due to the choice of dataset, and are in-line with previous works. For instance, Yang et al. (2020) obtains a top-1 (unweighted) accuracy of 56%, predicting from 143 classes on the People Subtree, whereas we obtain a top-1 (unweighted) accuracy of 50%, predicting from 284 classes.

The following tables detail additional dataset and data augmentation information.

	Training Split Size	Attribute-Labeled Training Size	Eval Size	Test Size
CelebA	162770	162770	19867	19962
ImageNet	124693	5861	5327	5327

	Normalization	Optimal Data Augmentation
CelebA	By mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225], center cropped and resized to (224, 224)	Random resized crop of (224, 224) from (256, 256) and random horizontal flips
ImageNet	By mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225], center cropped and resized to (224, 224)	Random resized crop of (224, 224) from (256, 256) and random horizontal flips

B.3 Experiment Hyperparameters

In this section, we detail the hyperparameters used in our experiments. All results provided are over three random seeds. Unless otherwise-specified, these hyperparameters were selected by a grid search over the hyperparameter candidate values listed in the below table.

	Range
Learning Rate	1e-2, 1e-3, 1e-4, 1e-5
Batch Size	32, 128
Weight Decay	1e-1, 1e-4, 0
Dropout	0, 0.5
Group learning rates (WA-GDRO)	1, 0.1, 0.01
Gradient penalty (IRM)	0.2, 1, 5
Groups sampled per batch (IRM)	1, 4, 16
Initial Epochs (UWA-JTT)	1, 5, 30
Importance weight (UWA-JTT)	1, 5, 20, 50

Hyperparameters were selected from the above choices to optimize reweighted accuracy/mAP. We note that there is an interesting relationship between hyperparameter tuning and the fairness of various learning algorithms, but this is beyond the scope of this work and we defer interested readers to Hooker (2021). Our methodology is aimed at mitigating the influence of hyperparameters such that our empirical findings reflect meaningful differences between methods rather than differences in hyperparameter tuning/specifications.

Table 4 is an experiment run on the CelebA-SL task. All hyperparameters seen are chosen to match Shrestha et al. (2021) as closely as possible. As with the original paper, we use a Resnet-18 (He et al., 2016) trained with SGD with momentum 0.9 and without data augmentation. The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
U-ERM	1e-3	128	0	0	50	N/A
W-ERM	1e-5	128	1e-1	0	50	N/A
WA-GDRO	1e-5	128	1e-1	0	50	Group learning rate of 0.01
IRM	1e-4	128	0	0	50	Gradient penalty of 1
U-SD	1e-4	128	1e-5	0	50	Per class $\lambda = (10, 10)$, $\gamma = (0.44, 0.25)$.

Table 5 is an experiment run on the CelebA-SL task. All hyperparameters seen are chosen to match Liu et al. (2021) as closely as possible. As with the original paper, we use a Resnet-50 (He et al., 2016) trained without data augmentation, trained with SGD with momentum 0.9, and pretrained on ImageNet. The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
U-ERM	1e-4	128	1e-4	0	50	N/A
W-ERM	1e-4	128	1e-4	0	50	N/A
WA-GDRO	1e-5	128	1e-1	0	50	Group learning rate of 0.01
UWA-JTT	1e-5	128	1e-1	0	50	Importance weight of $\lambda = 50$, using ERM model trained for 1 epoch.

Table 6 is an experiment run on the CelebA-ML task. All hyperparameters seen are chosen to match Wang et al. (2020) as closely as possible. As with the original paper, we use a Resnet-50 (He et al., 2016) trained with data augmentation and Adam, pretrained on ImageNet. The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
U-ERM	1e-4	128	0	0.5	50	N/A
W-ERM	1e-4	128	0	0.5	50	N/A
Ind	1e-4	128	0	0.5	50	N/A
Independent SP	1e-4	128	0	0.5	50	N/A
Disc	1e-4	128	0	0.5	50	N/A
A-Cens	1e-4	128	0	0.5	50	Training ratio (adversarial:main) 3:1, confusion loss weight = 1.0

The CelebA-ML task depicted in Figures 2, 3, 4, 6, 8, 9, and 7 are run with the below hyperparameters, determined by grid search. We also use a Resnet-50 He et al. (2016) trained with data augmentation and Adam and pretrained on ImageNet—the same settings as Wang et al. (2020).

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
U-ERM	1e-4	32	0	0.5	30	N/A
W-ERM	1e-4	32	0	0.5	30	N/A
Ind	1e-4	32	0	0.5	30	N/A
WA-GDRO	1e-4	32	0	0.5	30	Group learning rate of 0.1
UWA-JTT	1e-4	32	0	0.5	30	Importance weight $\lambda = 20$, using ERM model trained for 1 epoch

The CelebA-SL task depicted in Figures 2, 3, 4, 6, 8, 9, and 7 are run with the below hyperparameters, determined by grid search. We also use a Resnet-50 He et al. (2016) trained without data augmentation, with SGD with momentum 0.9, and pretrained on ImageNet—the same settings as Liu et al. (2021).

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
U-ERM	1e-4	128	1e-1	0	50	N/A
W-ERM	1e-5	128	1e-1	0	50	N/A
Ind	1e-4	128	1e-4	0	50	N/A
WA-GDRO	1e-4	128	1e-1	0	50	Group learning rate of 0.01
UWA-JTT	1e-4	32	1e-1	0	50	Importance weight $\lambda = 5$, using ERM model trained for 1 epoch
IRM	1e-4	32	1e-1	0	50	Gradient penalty of 1
Uniform IRM	1e-4	128	1e-1	0	50	Gradient penalty of 1, 16 groups sampled per batch
U-SD	1e-4	128	1e-4	0	50	Per class $\lambda = (10, 10)$, $\gamma = (0.44, 2.5)$

Tables 2, 3, and 7 are experiments run on the ImageNet dataset. Hyperparameters were chosen by grid search. We use a Resnet-50 He et al. (2016) trained with SGD with momentum 0.9, with standard ImageNet pretrained weights—note that the ImageNet subset used for pretraining does not intersect with the ImageNet People Subtree we train on.[1] The hyperparameters are listed below.

[1]Further note that while we initialize our network with standard ImageNet pretrained weights (trained on a different subset of ImageNet than we use), some of our experiments involve also pretraining on a subset of ImageNet that we do use (see Table 7).

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
U-ERM	1e-4	64	1e-4	0	100	N/A
W-ERM	1e-4	64	1e-2	0	50	N/A
W-SqrtERM	1e-4	64	1e-2	0	50	N/A
W-SqrtERM Distilled	1e-4	64	1e-2	0	50	Distill weight = 1.0
WA-GDRO	1e-4	64	1e-2	0	50	Group learning rate of 0.01
UWA-JTT	1e-5	64	1e-1	0	50	Importance weight of $\lambda = 5$, using ERM model trained for 5 epoch.
Ind	1e-4	64	1e-1	0	50	N/A
Ind Distilled	1e-4	64	1e-1	0	50	Distill weight = 1.0
U-SD	1e-5	64	1e-1	0	50	Per class $\lambda = 10$, $\gamma = 0$