Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers

Dominik Zietlow\textsuperscript{1,2,†,}, Michael Lohaus\textsuperscript{1,3,†}, Guha Balakrishnan\textsuperscript{4}, Matthäus Kleindessner\textsuperscript{1}, Francesco Locatello\textsuperscript{1}, Bernhard Schölkopf\textsuperscript{1,2}, Chris Russell\textsuperscript{1}

\textsuperscript{1}Amazon Web Services, Tübingen, \textsuperscript{2}MPI-IS Tübingen, Germany, \textsuperscript{3}University of Tübingen, Germany, \textsuperscript{4}Rice University, USA

Abstract

Algorithmic fairness is frequently motivated in terms of a trade-off in which overall performance is decreased so as to improve performance on disadvantaged groups where the algorithm would otherwise be less accurate. Contrary to this, we find that applying existing fairness approaches to computer vision improve fairness by degrading the performance of classifiers across all groups (with increased degradation on the best performing groups).

Extending the bias-variance decomposition for classification to fairness, we theoretically explain why the majority of fairness methods designed for low capacity models should not be used in settings involving high-capacity models, a scenario common to computer vision. We corroborate this analysis with extensive experimental support that shows many of the fairness heuristics used in computer vision also degrade performance on the most disadvantaged groups. Building on these insights, we propose an adaptive augmentation strategy that, uniquely, of all methods tested, improves performance for the disadvantaged groups.

1. Introduction

High-capacity neural classifiers achieve state-of-the-art performance on most computer vision tasks when assessed by overall test set accuracy. However, researchers have begun examining the unfairness of these models. Here, we use ‘unfairness’ to refer to systematic accuracy differences across protected subgroups defined by human-sensitive attributes like gender and race [7,8,36].\textsuperscript{1} These differences in accuracy can harm certain population groups, and as a result, numerous strategies for training models that match various measures of accuracy across subgroups have been developed [2,5,33,85]. Typically, such methods quantify unfairness by comparing accuracy-related rates between various groups, for example the Difference of Equal Opportunity (DEO) [36] compares groupwise true positive rates.

Many recent computer vision fairness studies are motivated by a fairness-accuracy trade-off originally targeting low-capacity models [1,2,15,22,47,56,85] where high accuracy on better performing (and often larger) groups comes at the cost of lower accuracy on the worse performing (and often smaller) groups. In this case, it is possible to increase fairness by reducing the accuracy on the best-performing group and increasing the accuracy on the worst-performing group, see models A and B in Figure 1.

We revisit this trade-off and show that it does not hold when using high-capacity neural classifiers prevalent in computer vision (see Figure 2). Instead, many fairness methods degrade the accuracy of networks on all groups, with a greater degradation occurring for the better performing groups. This increases fairness, but at the cost of producing a worse performing classifier (Figure 1 model C).

The phenomenon of balancing fairness by degrading the performance on the better off groups is referred to as leveling down in law and philosophy, where it has received substantial criticism [12,19,26,41,57]. The behavior we are concerned with is even more extreme than the typical levelling down – rather than just lower the performance on the best performing groups, every group is worse off.

If fairness methods decrease performance for all groups, they are Pareto Inefficient with respect to group accuracy and should not be used in contexts where the accuracy of any group is a primary concern.\textsuperscript{2} For example, we find that for classifiers on CelebA [52] regularized by a DEO fairness measure, the increased fairness comes at the cost of degraded performance for every group including the worst performing group (see Figure 2). We attribute this problem to two issues:

\textsuperscript{1}In this work, we focus on fairness measures that compare accuracy across groups. These are in contrast to measures such as demographic parity [14,28], which matches the proportion of positive decisions per group.

\textsuperscript{2}Various works in fairness have made use of Pareto Efficiency. It has been used both to refer to trading-off global accuracy against notions of fairness [76], and for the notion of trading-off per group accuracies against each other [56]. We only refer to the second case.
High-capacity classifiers fit training data nearly perfectly: Most methods within the fairness community are designed for low-dimensional data where a classifier cannot fit data from multiple distinct distributions well, even on training data [1, 2, 15, 22, 36, 47, 56, 85]. This is not the case in computer vision, where high-dimensional data and high-capacity models mean that near-zero training error is common [87]. Fairness notions that are aligned with the perfect classifier are therefore trivially satisfied on training data. Unfortunately, most existing methods do not take this into account and enforce fairness constraints on the training set [2, 10, 22, 24, 25, 27, 44, 47, 53, 56, 61, 82, 85].

Inappropriate evaluation of fairness methods: Most papers presenting fairness methods report a combination of accuracy and a fairness measure such as DEO, and take a decrease in accuracy and an improved fairness measure as an indication that the method is successfully trading off fairness against accuracy [2, 65, 66, 81, 82, 85]. The crucial question of whether the learned models work better for more disadvantaged groups remains unanswered. Consequently, the choice of metrics may mask a systematic deterioration of performance across all groups.

To address these limitations, we make three contributions:

1. We revisit the bias-variance trade-off in a simple decomposition of fairness on test data into training error (bias) and generalization error (variance), and observe that in situations where the training error goes to zero, any measure of fairness must be dominated by generalization error. As such, methods not using held-out data for fairness constraints cannot work for the high-capacity classifiers common to computer vision.

2. We perform an extensive evaluation of existing fairness methods and show that their reported improvement in fairness metrics is accompanied by worse performance across all groups.

3. To confirm our theoretical analysis that better generalization is key to improving performance on the most disadvantaged groups, we explore the use of data augmentation combined with adaptive sampling. We propose a novel GAN-based augmentation and show that it improves fairness by improving accuracy on the most disadvantaged groups on the CelebA dataset.

2. Related Work

Notions of fairness: We focus on notions of fairness that aim to balance classifier errors across population subgroups, for example, by matching error rates across genders or different racial groups. See [78] for a summary of other forms of fairness, and [79] for when such definitions are inappropriate. One common fairness measure is ‘equal opportunity’ (EO) [36]. This requires a classifier to have equal true positive rates (TPR) on each subgroup. We use $A$ to refer to the set of protected attributes, such as race or gender, $Y \in \{0, 1\}$ to refer to the ground-truth label of a data-point, and $\hat{Y} \in \{0, 1\}$ the prediction of $Y$. For two groups $a, a' \in A$ the violation of equal opportunity is measured by the difference in equal opportunity (DEO) defined as

$$|\mathbb{P}(\hat{Y} = 1 | Y = 1, A = a) - \mathbb{P}(\hat{Y} = 1 | Y = 1, A = a')|.$$  

(1)
We measure the difference in equalized odds (DEOdds) by

$$\sum_{y} \left[ \mathbb{P}(\hat{Y} = 1|Y = y, A = a) - \mathbb{P}(\hat{Y} = 1|Y = y, A = a') \right].$$

Another fairness notion of recent interest is min-max fairness [22, 55]. In contrast to EO (and the other group fairness notions) min-max fairness does not equalize a statistic across groups, but solely strives to decrease the classification error for the subgroup with the highest error as much as possible. Formally, min-max fairness methods optimize

$$\min_{\hat{Y}} \max_{a \in A} \mathbb{P}(\hat{Y} \neq Y|A = a).$$

As a measure of min-max fairness we report the minimum group accuracy (i.e., 1 – error) and similarly the minimum group TPR. Note that the minimum group TPR is trivially maximized by the constant classifier $\hat{Y} = 1$.

While many variants of these error-matching fairness definitions exist\(^3\), the motivation for all measures is broadly the same: to prevent a tyranny of the majority, in which accuracy on smaller groups of people are neglected to achieve higher accuracy on larger groups.

**Fairness in computer vision:** Following foundational work [13] that identified the systematic failings of face analysis systems on particular racial and gender demographics, algorithmic fairness has attracted growing attention in the computer vision community. Image datasets are known to be biased due to sampling inequalities [3, 64, 75], and human face datasets have been particularly scrutinized [6, 45, 49, 50, 58] as models trained on these data can exhibit systematic failings with respect to attributes protected by the law [46]. Multiple approaches to mitigating dataset bias include collecting more diverse examples [58], using image synthesis to compensate for distribution gaps [6, 50, 72, 74, 82], and resampling [51]. More recently, approaches have been proposed to mitigate the systematic biases of deep learning-based vision models [18, 30, 34, 40, 44, 51, 65, 66, 80, 81].

Outside of computer vision, the most common approaches add additional fairness measures to loss functions [9, 47, 53, 61, 69, 83], enforcing fair representations that are independent of protected attributes [10, 29, 54, 86], and augmenting training data to promote balance [66]. Most of these studies assume an accuracy-fairness tradeoff which we show can lead to suboptimal training or misleading evaluation for deep neural networks.

**Active sampling for min-max fairness:** While the methods of [22, 55] achieve min-max fairness by means of reweighting and retraining, the recent method of [1] uses adaptive sampling and standard SGD updates to minimize a differentiable proxy of (3). This makes the latter method easily applicable to deep neural network training.

### 3. On Accuracy-based Fairness in Low- and High-capacity Classifiers

This section outlines the challenges peculiar to enforcing fairness in the high-capacity classifiers common to computer vision. In particular, we focus on accuracy-based notions of fairness that aim to match accuracy across groups.

As noted by Wachter et al. [79], any accuracy-based fairness measure\(^4\) is trivially satisfied by a classifier with zero error. However, as the typical datasets (see [2, 43, 85] for a range of examples) used in the fairness literature are low-dimensional, and with large amounts of label volatility (see Section 3.1), classifiers with zero error do not occur in practice, even on the training data. Training a low-capacity classifier with accuracy-based fairness constraints on the training set is a common approach. Where training error remains high, such approaches can remain effective even if fairness constraints are only enforced over the training data.

This is not the case when enforcing fairness of deep-learning based classifiers on computer vision datasets. Such datasets are empirically shatterable [77]. Even if the images in the training set are randomly relabelled, it is possible to learn a classifier with zero error on the training set [87]. In such scenarios, accuracy-based fairness definitions are trivially satisfied over the training set.

In the following subsection, we formalize the gap between fairness on training and held-out data and discuss its implications for computer vision.

#### 3.1. Bias-variance decomposition for classification

It is common in statistics to decompose the error into three terms: irreducible label noise $N$, a bias $B$ representing how well the regressor can fit the dataset, and a variance $V$ representing additional error induced when generalizing to new data. The standard decomposition was formalized around the squared loss [31], extended to the zero-one loss [48] and generalized to arbitrary losses [23].

We build on the latter formulation, which we summarize in a condensed manner—see [23] for more details. We consider the task of learning $f \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$, where $f(x)$ should be a good prediction of the label $y$ of input point $x$. The quality of a prediction is measured by a loss function $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$. The optimal prediction for input $x$ is

$$y_*(x) = \arg \min_{y'} \mathbb{E}_{y|x} [L(y, y')]$$

and the expected loss of model $f$ on $x$ is $\mathbb{E}_{y|x} [L(y, f(x))]$. The conditional expectation corresponds to the fact that,

\(^3\)As a starting point, [79] identified 17 accuracy-based measures excluding min-max fairness.

\(^4\)Wachter et al. refer to these accuracy-based measures as bias-preserving fairness metrics.
in general, the label $y$ is a non-deterministic function of input $x$. We learn the model $f$ based on a training set $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and in order to remove the dependency on $D_n$ we consider $\mathbb{E}_{D_n, y|f} \left[ L(y, f(x)) \right]$ as the expected error of a classifier trained over both groups rather than the error of a classifier trained on a fixed dataset, the bias $B(x)$ (error induced by systematic model imperfection), and the variance $V(x)$ (error difference from the main prediction):

$$
B(x) = L(y_1(x), y_m(x)), 
$$

$$
V(x) = \mathbb{E}_{D_n} \left[ L(y_m(x), f(x)) \right], 
$$

$$
N(x) = \mathbb{E}_{y|x} \left[ L(y, y_1(x)) \right].
$$

As [23] shows, for certain loss functions $L$ (including but not limited to squared, zero-one loss, and false negative rate)\textsuperscript{5} we can decompose

$$
\text{err}_x := \mathbb{E}_{D_n, y|x} \left[ L(y, f(x)) \right] = c_1(x)N(x) + B(x) + c_2(x)V(x)
$$

for some $c_1(x), c_2(x) \in \mathbb{R}$.

### 3.2. Expected fairness violations

We make the simplifying assumption that we want to enforce error parity across only two groups $A$ and $B$. If we define notation for both groups, we use $G$ in place of $A$ or $B$. We depart slightly from [23] as we consider the per-group error of a classifier trained over both groups rather than the classifier’s overall error. The expected fairness violation for the two-groups case $E_{\text{fair}}$ can be defined as

$$
E_{\text{fair}} = \left| \mathbb{E}_{x \in A} \left[ \text{err}_x \right] - \mathbb{E}_{x \in B} \left[ \text{err}_x \right] \right|. 
$$

With auxiliary definitions

$$
B_G = \mathbb{E}_{x \in G} \left[ B(X) \right],
$$

$$
V_G = \mathbb{E}_{x \in G} \left[ c_2(x) V(X) \right],
$$

$$
N_G = \mathbb{E}_{x \in G} \left[ c_1(x) N(X) \right],
$$

the fairness violation can be rephrased as

$$
E_{\text{fair}} = \left| N_G + B_A - N_B - B_B \right|. 
$$

\textsuperscript{5} The use of zero-one loss allows us to consider the bias-variance decomposition of DEOdds, while false negative rate allows us to consider DEO.

### 3.2.1 Fairness for low-capacity classifiers

For low-capacity models, the variances are strongly dominated by the noises and biases [37]. That is $N_G + B_G \gg V_G$ and the fairness violation can be approximated as

$$
E_{\text{fair}} \approx |N_A + B_A - N_B - B_B|. 
$$

Importantly, $N_G + B_G$ can be directly estimated from the per group training set, and as such methods that match the loss across different groups at training time are likely to work here. Indeed, many methods [2, 36, 85] enforce fairness by explicitly balancing errors on the training set, under the reasonable assumption that these fairness constraints will generalize well to unseen data. The work [25] addressed this by minimizing a combination of unfairness on the training set and an upper bound of the generalization error. However, in practice, the majority of approaches are effective without considering generalization error.

### 3.2.2 Fairness for high-capacity classifiers

In contrast, for classification tasks in computer vision, the typical behavior is very different. While human labelling of image data is often a noisy process [59], the majority of computer vision datasets do not model label noise and often either only collect one label per datapoint [52], or explicitly denoise the collected labels [21] and only use the most common label per datapoint as the true datapoint.\textsuperscript{6} As such, when we treat dataset acquisition as a stochastic process that can only assign one label per datapoint, eq. (6), which measures the expected loss caused by disagreement between labels, is zero. This point is key to our argument; where multiple disagreeing labels are collected for individual datapoints, it may be possible for methods to improve these forms of fairness without making use of held-out data.

For a computer vision classifier trained to convergence on a fixed dataset, the bias terms also disappear [84]. Here, data is high-dimensional and models have essentially arbitrarily high capacity. Hence, the perfect classifier, which has zero error and predicts the ground-truth label for each datapoint, lies in $\mathcal{F}$. Hence, $V_G \gg B_G \approx 0$, and the fairness violation is dominated by the generalization error

$$
E_{\text{fair}} \approx |V_A - V_B|,
$$

\textsuperscript{6} i.e., given an image labeled as cat twice, and dog once, the image is simply treated as being labelled cat.
Given this, it is unsurprising that these methods for enforcing fairness on a training set are not popular in computer vision, and instead research typically centers around heuristics, such as reweighting samples [51], training multiple classifiers on data subsets, and averaging them [70, 71], or data augmentation [50]. With such a wide range of approaches, it is not possible to formally analyze them, beyond saying that they do not use held-out validation data to estimate generalization error, but they all successfully trade-off accuracy against a wide range of fairness metrics.

3.2.3 Rethinking fairness measures in computer vision

We have shown it is not possible to predict per-group error rates on the test set without computing fairness measures on held-out data. Yet the vast majority of methods do not do so, and as demonstrated in Section 5, still consistently improve fairness according to standard metrics such as equalized odds or equal opportunity. How is this possible?

One answer lies in the definitions of equalized opportunity (1) and odds (2) – they are perfectly satisfied by random or constant classifiers. As such, it is possible to increase fairness by decreasing performance on all groups, rather than by rebalancing the errors from one group to another.

Empirically, we find that this is exactly what occurs in practice: improvements to DEO and DEOdds are accompanied by a degradation in the accuracy for all groups, and not a rebalancing of accuracy from one group to another (see Figures 2, 4). While it is not possible to predict the per-group error for unseen data without making use of held-out data, it is always possible to degrade performance by injecting noise into gradients (which may be happening with regularized approaches [61, 83]), augmenting datasets with inappropriate synthetic examples (which might be happening with [66]), or by a wide range of well-motivated and sensible-sounding heuristics, including early stopping.

In light of this issue, when evaluating accuracy-based model fairness, we recommend choosing a fairness metric that explicitly requires improving performance on the disadvantaged groups. In our experiments, we measure the accuracy of the worst performing group, i.e., min-max fairness of eq. (3). Another natural choice would be the true positive rate on the worst performing group, but this is easily gamed by classifiers with high false positive rates.

4. Improving accuracy on disadvantaged groups with synthetic data

In the previous section, we presented the importance of evaluating fairness based on the worst-performing group’s accuracy. Next, we consider how to use this insight, along with the result in (15) to better incorporate fairness during the model training. Expression (15) shows that fairness violations in high-capacity models are dominated by the variance of the bias-variance decomposition. Consequently, one way to improve fairness without decreasing the performance for all groups is to decrease the variance of the worst-off group. It is well-known, and experimentally verified [11], that variance typically decreases as the training set increases in size and diversity. Given a fixed dataset, data augmentation is the only way to achieve that. Taking these insights into account, we propose an adaptive augmentation strategy that increases the diversity of samples for the worst-performing group in a classification task.

There are three technical challenges that we must address with our method. How to: (i) decide which group(s) requires augmentation (ii) generate high fidelity in distribution data, and (iii) reliably condition the augmented data on the protected groups and automatically provide target labels. We tackle these challenges by: (i) deploying adaptive sampling strategies using held-out data to prioritize worst-performing groups, (ii) using invertible GAN architectures and latent space traversals to edit images, and (iii) proposing g-SMOTE, a generalization of the synthetic minority oversampling technique (SMOTE) [16], which produces new labeled training images by traversing GAN latent space.

4.1. Adaptive sampling

At each training iteration, we sample a random batch from one of two datasets: the original training dataset, and an extended training dataset. The extended training set is equal to the original training set at initialization. After each iteration, we determine the worst-performing group using a held-out evaluation dataset, augment a random batch from this group (using g-SMOTE in next section), and add it to the extended set. This approach is inspired by past studies [67, 68] that evaluate the generalization when training on partially augmented data.

We select a batch from the original dataset with probability $\lambda$ (and $1 - \lambda$ from the extended dataset). The parameter $\lambda$ allows us to maintain a certain proportion of the original data in the training set, preventing the extended training set from becoming predominantly augmented data. Note that this is more involved than simply balancing the number of elements in each group. Differences in generalization performance can stem from strongly imbalanced group sizes, but also from characteristics of either group that make generalization intrinsically harder.

4.2. Generalized SMOTE: g-SMOTE

Given images from the original dataset, we need a procedure to generative new synthetic images along with attribute labels. To accomplish this, we combine SMOTE [16] with the rich representational and generative capabilities of modern GANs. SMOTE is a simple sampling strategy to overcome imbalanced data that is based on linear feature interpolation between a datapoint and a random datapoint within
Algorithm 1 Adaptive Sampling

1: Inputs:
   Hyper-parameter $\lambda \in [0, 1]$
   Train dataset $D_{\text{train}} = \{(x_i, y_i), (x_1, y_1), \ldots \}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}$
   Evaluation dataset $D_{\text{eval}} = \{(x^e_i, y^e_i), (x^e_1, y^e_1), \ldots \}, x^e_i \in \mathcal{X}, y^e_i \in \mathcal{Y}$
   Classifier $c_{\phi}: \mathcal{X} \rightarrow \mathcal{Y}$ (parameterized by $\phi$)
2: Initialize: $D_{\text{Aug}} := D_{\text{train}}$
3: for $t = 1, \ldots, 100$ training steps do
4:   With probability $\lambda$, uniformly sample $(x_i, y_i) \in D_{\text{train}}$, otherwise sample $(x_i, y_i) \in D_{\text{Aug}}$
5:   Update $\phi$ according to learning objective
6:   Determine weakest group based on learning objective and $D_{\text{eval}}$ and augment corresponding $x_{\text{aug}}, y_{\text{aug}}$ from that group
7:   $D_{\text{Aug}} \leftarrow D_{\text{Aug}} \cup \{ (x_{\text{aug}}, y_{\text{aug}}) \}$
8: end for

the set of its $k$ nearest neighbors. It has inspired many other approaches [17, 35, 38, 88]. Modern invertible GAN architectures [24, 27, 32, 63] allow one to ‘embed’ images into their latent spaces, which are known to be particularly effective domains to compare and edit images [6, 73]. We propose to use SMOTE in the GAN latent space to generate new diverse synthetic images along with attribute labels.

We propose a generalized variant of SMOTE (g-SMOTE), that extends the classical SMOTE strategy – linear interpolation between a datapoint and a random point among the datapoints $m$ nearest neighbors – to uniform sampling within a $k$-dimensional simplex formed by $k$ of the $m$ nearest neighbors, to improve data diversity. Given a datapoint and its $m$ nearest neighbors with the same target attribute, we randomly choose $k$ neighbors that span a simplex of dimension $k - 1$ or smaller in the much higher dimensional GAN latent space. Within this simplex, we uniformly sample points in latent space and use the GAN generator to render their images. Our key assumption is that the simplex covers a label-consistent volume in latent space, i.e., that every image stemming from that region shares the same target label. Algorithm S2 describes the data augmentation procedure, Figure S2 illustrates the interpolation mechanism, and Figure 3 shows a real-world example for $k = 3$. See supplementary materials for the effects of varying $k$ and a comparison of SMOTE and g-SMOTE.

5. Experiments

We show that a wide range of methods for fairness in computer vision decrease performance on the worst group despite reporting improved fairness, and that, in contrast, our approach improves performance. As such, we follow existing experimental protocols as closely as possible. We show results on CelebA where all methods are reported to work, we use existing code, and compare each method to its corresponding unfair baseline. We demonstrate that:

Common fairness approaches in computer vision reduce performance for protected groups compared to unfair baselines. As such, we evaluate methods with respect to their min-max performance on accuracy and TPR.

Adaptive debiasing with g-SMOTE improves min-max performance. See Section 5.1, and Figure 4 for experimental results. The benefit of adaptive debiasing is emphasized when training on a subset of data, because the GAN allows for effective unsupervised data augmentation.

Adaptive debiasing with g-SMOTE works with cross-sectional groups of multiple protected attributes. See Section 5.2 for experimental results. We show our method works with the four groups defined by the cross product of gender and target attribute binary groups.

g-SMOTE produces better data diversity than popular augmentation strategies. See Section 5.3. When combining g-SMOTE augmentation with popular augmentation methods, such as random flips, crops, rotations and RandAugment, we always observe an improvement in accuracy.

5.1. Evaluating min-max fairness on existing methods and adaptive debiasing

Our experimental setup follows [66] and [82] with full details in Appendix D. We compare against methods to improve fairness, including oversampling, domain discriminative training [70, 71], domain independent models [82], an adversarial approach [4], regularization [61, 83], FairMixup [20] and GAN-based offline dataset debiasing [66].

We summarize results in Figure 4. Some methods use a single multitask model (annotated with *), others use multiple single task models. Note that FairMixup uses a higher resolution (256 px) and a lower capacity model (ResNet18), which leads to a substantially different baseline performance. The only methods to increase the performance on the less accurate groups are g-SMOTE with
Adaptive Sampling

Max. Group Acc.

90.5

93.0

93.5

94.0

94.5

GAN Debiasing

Regularized


Domain Indep.

Weighting

GAN Debiasing

Domain Disc.

g-SMOTE

g-SMOTE + Adaptive Sampling

Figure 4. Fairness methods on the CelebA dataset. We report mean scores over 13 labels [66] call gender independent. Left: Change in minimum group accuracy difference relative to the unconstrained classifier. Only our approaches g-SMOTE and Adaptive g-SMOTE improve accuracy on the worst performing groups. All other methods lower the accuracy. Top Right: A decomposition of the change in accuracy into overall change, change in the best performing group, and the worst. Center Right: Decomposition of the change in True Positive Rate (TPR). Despite DEO being defined in terms of TPR, all but one of the fairness method decrease the TPR. Note that our g-SMOTE based approaches improve accuracy at the cost of decreasing TPR. Bottom Right: plots of DEO and DEOdds. These standard fairness measures track each other closely, with a high score in one corresponding to a high-score in the other. However, they are unpredictable of improvements in TPR or accuracy on any groups. Methods marked with * are based on the code-base of [82], training one model for multi-task classification. Other methods implement one-task classification models.

Figure 5. A comparison of fairness methods and unmodified classifiers showing the trade-off behaviour between better and worse performing groups. Results are averaged over 13 labels with gender independent quality. The shaded rectangles indicate Pareto-efficient areas in which neither of the groups gained performance over the unmodified model. Note that FairMixup uses a smaller model and higher resolution. Our methods are the only ones to improve worst group performance over the relevant baseline.

5.2. Adaptive g-SMOTE on multiple groups

Adaptive g-SMOTE is trivially extendable to multiple protected groups. We compare the adaptive g-SMOTE sampling with the regular g-SMOTE augmentation for two protected groups (“male” / not “male”) and four protected groups (“male” / not “male”, target attribute / not target attribute) in Table 1 and Table 2 respectively. The 32 target labels are selected such that each of the four protected groups has at least 0.3 % of the total number of datapoints.

This comparison leads to three interesting findings: (1)

Table 1. Sampling strategies on CelebA. We report the means over 32 labels. The protected attribute is “male”. The models shown have the highest min. group accuracy over the training period.

and without adaptive sampling. The baselines lower accuracy across both protected groups. For the offline debiasing approach, we used our own implementation and the same GAN model as used for the g-SMOTE methods.

The argument set out in Section 3.1 suggests that for high capacity models, balancing datasets with adaptive sampling is less important than improving generalization by generating diverse synthetic data. Empirically, we find this to be the case. Using g-SMOTE without adaptive sampling, we find that performance can be improved for all groups, with only a small decrease in min-group accuracy compared to adaptive sampling, see Table 1. We report the min. group accuracy per attribute in Table S2.

<table>
<thead>
<tr>
<th>2 Protected Groups</th>
<th>Baseline</th>
<th>Adaptive g-SMOTE</th>
<th>g-SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10k train images</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc.</td>
<td>89.41</td>
<td>89.57</td>
<td>89.58</td>
</tr>
<tr>
<td>Min. grp. acc.</td>
<td>85.68</td>
<td>85.93</td>
<td>85.91</td>
</tr>
<tr>
<td>DEO</td>
<td>23.18</td>
<td>25.00</td>
<td>24.65</td>
</tr>
<tr>
<td>DEOdds</td>
<td>33.63</td>
<td>35.51</td>
<td>34.64</td>
</tr>
</tbody>
</table>


adaptive sampling gives a substantial improvement in min-group accuracy as well as improving the fairness measures.

Based on our theoretical analysis and experimental findings, we make two key recommendations for practitioners:

1. Evaluate a model using the error of the worst performing group: If accuracy-based notions of fairness are appropriate for a given scenario, i.e., particular groups are being disadvantaged by a high error rate, a better measure for fairness than the difference of error rates is the error over the worst performing group. If the rate is sufficiently low, the method can be safely deployed, otherwise not.

2. Gather more data for the worst performing groups: We have shown that the key issue preventing fairness in computer vision is an inability to generalize well. However, outside of standardized benchmarks with fixed training sets, generalization can be improved by gathering more data. This leads to a direct approach where one iteratively evaluates on held-out data and gathers additional data on the worst performing group. This is equivalent to algorithm 1 of [1], and under certain constraints is guaranteed to minimize the worst error over all groups. Other works [13] have called for more diverse datasets, but approaching this as a min-max problem tells us how to grow the dataset, and why diversity improves performance even when it leads to a distribution mismatch between training and evaluation data.

Limitations: The analysis set out in Section 3 only holds for the accuracy-based notions of fairness that are satisfied by perfect classifiers. It may be possible to enforce other notions of fairness, such as demographic parity, on maximally accurate classifiers without held-out data. Moreover, a wide range of machine learning scenarios exist for which accuracy-based fairness is inappropriate [79].

Conclusion: The takeaway message from this work should not be that accuracy-based notions of fairness do not work in computer vision, but that if we measure and optimize the right thing, they can be made to work. Our analysis in Section 3 makes it clear that fairness on unseen data is primarily a problem of generalization. As such, three promising directions for improving performance in the most disadvantaged groups lie in Hyperparameter Optimization (HO), Network Architecture Search, and data augmentation, with HO already being used in fairness [42, 62, 76]. We have shown that data augmentation can improve performance on the most disadvantaged groups, but the use of these other techniques to improve worst group performance is a promising direction for future work.

Acknowledgments

CR is a member of the Governance of Emerging Technologies (GET) research programme at the Oxford Internet Institute. He is grateful to members of GET and the Trustworthy Auditing for AI project for informed discussion that helped shape this work.

<table>
<thead>
<tr>
<th>4 Protected Groups</th>
<th>Baseline</th>
<th>Adaptive g-SMOTE</th>
<th>g-SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10k train images</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc.</td>
<td>85.65</td>
<td>84.91</td>
<td>85.16</td>
</tr>
<tr>
<td>Min. grp. acc.</td>
<td>62.16</td>
<td>64.36</td>
<td>63.54</td>
</tr>
<tr>
<td>160k train images</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc.</td>
<td>87.37</td>
<td>85.77</td>
<td>87.27</td>
</tr>
<tr>
<td>Min. grp. acc.</td>
<td>61.74</td>
<td>68.06</td>
<td>61.84</td>
</tr>
<tr>
<td>DEOd</td>
<td>21.99</td>
<td>12.27</td>
<td>21.91</td>
</tr>
<tr>
<td>DEOd2</td>
<td>30.89</td>
<td>19.29</td>
<td>31.24</td>
</tr>
</tbody>
</table>

Table 2. Different sampling strategies on CelebA. Reported are the means over 32 labels. The protected attributes are “male” and the target attribute. The models shown have the highest min. group accuracy over the training period. DEO and DEOdds are only evaluated with respect to the attribute “male”.

6. Discussion

The majority of fairness methods examined report improved fairness measures of DEO and DEOdds despite decreasing both accuracy and TPR rates of the most disadvantaged groups. Nonetheless, we have shown that it is possible to improve performance for disadvantaged groups by the targeted generation of synthetic data. For the simple case of these high-capacity models (Table 1), in which protected groups correspond to gender, there is no apparent trade-off, and no need to target particular groups with adaptive sampling. Instead, we can uniformly generate samples and improve performance for everyone. In scenarios where we want high accuracy for groups corresponding to a combination of protected attribute and true label (e.g., males receiving a negative decision) this trade-off still exists. Here,
References


[22] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2021. 1, 2, 3


A. Additional Experimental Results

An extensive overview for the results visualized in Figure 4 are listed in Table S1.

B. GAN based SMOTE

Algorithm S2 GAN based synthetic minority oversampling

1: Inputs:
Hyper-parameters \( m, k \in \mathbb{N}^+ \)
Train dataset \( D_{\text{train}} = \{(x_0, y_0), (x_1, y_1), \ldots \}, x_i \in \mathcal{X}, y_i \in \mathcal{Y} \)
Trained invertible GAN \( E: \mathcal{X} \rightarrow Z, G: Z \rightarrow \mathcal{X}, Z = \mathbb{R}^n \)
datapoint \( (x_i, y_i) \)
2: Encode dataset
\( D_{\text{train}} \leftarrow \{(x_0, y_0), (x_1, y_1), \ldots \}, z_i = E(x_i) \)
3: Determine \( m \) nearest neighbours
\( N \leftarrow \{z_0, z_1, \ldots , z_N\} \)
\( \) according to \( z_i \) = \( \min_{z_j \in \{z_0, z_1, \ldots , z_N\}} \|z_i - z_j\|^2 \) \( s.t. \) \( x_i \neq y_j \)
4: Randomly choose \( k \) neighbours from that set
\( N \leftarrow \{z_0^{(1)}, z_1^{(1)}, \ldots , z_N^{(1)}\} \)
5: Uniformly sample \( z_{\text{augment}} \sim \mathcal{U}[\text{Simplex}[N]] \)
6: Return \( (G(z_{\text{augment}}), y_i) \)

C. Efficient Uniform sampling from the Convex Hull of Points

Here, we briefly summarize the strategy for efficiently uniformly sampling from the convex hull of a set of \( k \) points in an \( n \)-dimensional space where \( k \leq n + 1 \), as we have only been able to find related approaches in the literature for \( k = 3 \) [60].

Under minimal assumptions, e.g., the points are sampled with continuous noise, with probability 1 the \( k \) points form a \( k - 1 \) dimensional simplex, or generalized triangle. This allows us to sample efficiently using an inductive process. Given a point \( p_i \) sampled from a \( i \) dimensional simplex, \( S_i \) we can uniformly sample from an \( i+1 \) dimensional simplex, 

\( S_{i+1} = S_i \cup \{p_i\} \) by choosing

\( p_{i+1} = \lambda^{1/i} p_i + (1 - \lambda^{1/i})p_{i+1} \)  \( (S1) \)

where \( \lambda \) is sampled from the uniform distribution \( U[0, 1] \).

The base case is \( S_1 = \{p_1\} \), and \( p_1 = p_1 \).

Note that when the assumption fails, and points are colinear and do not form a simplex, the algorithm degrades reasonably. Sampled points will lie inside the convex hull of points, but the samples will not be uniform.

D. Architecture, Dataset and Details

The classifiers are built upon pretrained ResNet50 models [39]. For each of the binary attributes of the CelebA dataset, one classifier is trained. As the protected attribute we chose “male”\(^7\). Each model was trained for \( 3 \cdot 10^6 \) images and evaluated every 500 batches using Adam \( (lr = 10^{-3}) \) and a batch size of 64. Images were center-cropped and down-scaled to 128 \( \times \) 128. We use RandAugment with \( N = 3, M = 15 \) for every experiment unless otherwise stated. The reported numbers are the retrieved peak performance during the training period, evaluated on the hold-out evaluation dataset. We rely on the analysis of [66] and report the means over the attributes with gender independent label quality: “bags under eyes”, “bangs”, “black hair”, “blond hair”, “brown hair”, “chubby”, “eyeglasses”, “gray hair”, “high cheekbones”, “mouth slightly open”, “narrow eyes”, “smiling”, “wearing hat”. Rows labelled with * show results achieved using the codebase of [82]. For the invertible GAN model, we choose InvGAN [32], however the requirements on the model are

\(^7\)The labels in CelebA refer to an externally assigned perceived binary gender, and not to self-assigned gender identity. Although the binary nature of the label does not reflect the true distribution of either, we are restricted to the annotations available in the dataset.
Table S1. Fairness methods on the CelebA dataset. We report mean scores over the 13 labels [66] call gender independent. We report the model with the greatest min. group accuracy over the training period.

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Baseline</th>
<th>AdaptiveSMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavy Hair</td>
<td>76.95 ± 0.32</td>
<td>77.86 ± 0.11</td>
</tr>
<tr>
<td>Big Lips</td>
<td>80.49 ± 0.15</td>
<td>80.30 ± 0.16</td>
</tr>
<tr>
<td>Eyeglasses</td>
<td>99.23 ± 0.01</td>
<td>99.25 ± 0.02</td>
</tr>
<tr>
<td>Attractive</td>
<td>78.50 ± 0.14</td>
<td>79.17 ± 0.10</td>
</tr>
<tr>
<td>Brown Hair</td>
<td>81.26 ± 0.05</td>
<td>81.18 ± 0.12</td>
</tr>
<tr>
<td>Wearing Necklace</td>
<td>80.30 ± 0.01</td>
<td>80.30 ± 0.01</td>
</tr>
<tr>
<td>High Cheekbones</td>
<td>85.53 ± 0.11</td>
<td>85.61 ± 0.22</td>
</tr>
<tr>
<td>Receding Hairline</td>
<td>91.08 ± 0.09</td>
<td>91.40 ± 0.11</td>
</tr>
<tr>
<td>Wearing Hat</td>
<td>98.21 ± 0.02</td>
<td>98.22 ± 0.08</td>
</tr>
<tr>
<td>Black Hair</td>
<td>86.40 ± 0.10</td>
<td>87.10 ± 0.26</td>
</tr>
<tr>
<td>Gray Hair</td>
<td>95.28 ± 0.09</td>
<td>95.39 ± 0.08</td>
</tr>
<tr>
<td>Pale Skin</td>
<td>95.26 ± 0.06</td>
<td>95.43 ± 0.07</td>
</tr>
<tr>
<td>Smiling</td>
<td>91.64 ± 0.17</td>
<td>91.91 ± 0.09</td>
</tr>
<tr>
<td>Chubby</td>
<td>89.29 ± 0.02</td>
<td>89.35 ± 0.12</td>
</tr>
<tr>
<td>Young</td>
<td>82.28 ± 0.10</td>
<td>82.99 ± 0.24</td>
</tr>
<tr>
<td>Wearing Earrings</td>
<td>83.46 ± 0.14</td>
<td>83.67 ± 0.23</td>
</tr>
<tr>
<td>Big Nose</td>
<td>70.71 ± 0.29</td>
<td>71.12 ± 0.32</td>
</tr>
<tr>
<td>Oval Face</td>
<td>70.67 ± 0.10</td>
<td>71.15 ± 0.03</td>
</tr>
<tr>
<td>Bags Under Eyes</td>
<td>73.24 ± 0.37</td>
<td>73.69 ± 0.19</td>
</tr>
<tr>
<td>Bushy Eyebrows</td>
<td>87.08 ± 0.09</td>
<td>87.36 ± 0.08</td>
</tr>
<tr>
<td>Mouth Slightly Open</td>
<td>93.46 ± 0.12</td>
<td>93.58 ± 0.06</td>
</tr>
<tr>
<td>Rosy Cheeks</td>
<td>90.87 ± 0.11</td>
<td>90.97 ± 0.09</td>
</tr>
<tr>
<td>Arched Eyebrows</td>
<td>76.67 ± 0.35</td>
<td>76.83 ± 0.26</td>
</tr>
<tr>
<td>Blurry</td>
<td>95.58 ± 0.05</td>
<td>95.61 ± 0.06</td>
</tr>
<tr>
<td>Wearing Lipstick</td>
<td>86.14 ± 0.17</td>
<td>86.44 ± 0.14</td>
</tr>
<tr>
<td>Blond Hair</td>
<td>91.96 ± 0.05</td>
<td>92.10 ± 0.13</td>
</tr>
<tr>
<td>Heavy Makeup</td>
<td>84.13 ± 0.36</td>
<td>84.13 ± 0.01</td>
</tr>
<tr>
<td>Pointy Nose</td>
<td>69.20 ± 0.16</td>
<td>69.86 ± 0.11</td>
</tr>
<tr>
<td>Straight Hair</td>
<td>77.72 ± 0.42</td>
<td>77.98 ± 0.12</td>
</tr>
<tr>
<td>Bangs</td>
<td>94.67 ± 0.17</td>
<td>94.72 ± 0.04</td>
</tr>
<tr>
<td>Double Chin</td>
<td>91.43 ± 0.04</td>
<td>91.57 ± 0.17</td>
</tr>
<tr>
<td>Narrow Eyes</td>
<td>91.97 ± 0.08</td>
<td>92.32 ± 0.21</td>
</tr>
</tbody>
</table>

Table S2. Min. group accuracy for individual attributes of CelebA. Reported are the means and standard deviations over 3 restarts of all attributes with at least 11 positive and negative datapoints per group. The protected attribute is “male”. We evaluate the model with greatest min. group accuracy over the training period. Both methods were trained on 10000 images from the CelebA training set.

Regularized Approach. The regularized model presented in Figure 4 is based on a regularizer used in [83]. Given training data \( \mathcal{D} = \{x_i, y_i, a_i\}_{i=1}^n \) and a model \( f \) (as above) that outputs classification scores in \([0, 1] \), we define the regularizer \( R^{DEO} \) as

\[
R^{DEO}(f) := \left( \frac{1}{n_{11}} \sum_{D_{11}} f(x_i) - \frac{1}{n_{10}} \sum_{D_{10}} f(x_i) \right)^2,
\]

where \( D_{ga} = \{(\tilde{x}, \tilde{y}, \tilde{a}) : \tilde{a} = a, \tilde{y} = y, (\tilde{x}, \tilde{y}, \tilde{a}) \in \mathcal{D}\} \) and \( n_{ga} = |D_{ga}| \) for \( g \in \{0, 1\} \).

We simply add the fairness regularizer to the loss and trade-off fairness with the accuracy of the classifier via a hyperparameter \( \lambda \). We minimize \( \hat{L}(f) = \sum_{i=1}^n l(f(x_i), y_i) + \lambda R^{DEO}(f) \).