Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova,
Prem Nair, Kenji Hata, Olga Russakovsky
Princeton University
{zeyuwang, olgarus}@cs.princeton.edu

Abstract

Computer vision models learn to perform a task by capturing relevant statistics from training data. It has been shown that models learn spurious age, gender, and race correlations when trained for seemingly unrelated tasks like activity recognition or image captioning. Various mitigation techniques have been presented to prevent models from utilizing or learning such biases. However, there has been little systematic comparison between these techniques. We design a simple but surprisingly effective visual recognition benchmark for studying bias mitigation. Using this benchmark, we provide a thorough analysis of a wide range of techniques. We highlight the shortcomings of popular adversarial training approaches for bias mitigation, propose a simple but similarly effective alternative to the inference-time Reducing Bias Amplification method of Zhao et al., and design a domain-independent training technique that outperforms all other methods. Finally, we validate our findings on the attribute classification task in the CelebA dataset, where attribute presence is known to be correlated with the gender of people in the image, and demonstrate that the proposed technique is effective at mitigating real-world gender bias.

1. Introduction

Computer vision models learn to perform a task by capturing relevant statistics from training data. These statistics range from low-level information about color or composition (zebras are black-and-white, chairs have legs) to contextual or societal cues (basketball players often wear jerseys, programmers are often male). Capturing these statistical correlations is helpful for the task at hand: chairs without legs are rare and programmers who are not male are rare, so capturing these dominant features will yield high accuracy on the target task of recognizing chairs or programmers. However, as computer vision systems are deployed at scale and in a variety of settings, especially where the initial training data and the final end task may be mismatched, it becomes increasingly important to both *identify* and develop strategies for *manipulating* the information learned by the model.

Societal Context. To motivate the work of this paper, consider one such example of social bias propagation: AI models that have learned to correlate activities with gender [4, 7, 52, 2]. Some real-world activities are more commonly performed by women and others by men. This real-world gender distribution skew becomes part of the data that trains models to recognize or reason about these activities.[1] Naturally, these models then learn discriminative cues which include the gender of the actors. In fact, the gender correlation may even become *amplified* in the model, as Zhao et al. [52] demonstrates. We refer the reader to e.g., [34] for a deeper look at these issues and their impact.

Study Objectives and Contributions. In this work, we set out to provide an in-depth look at this problem of training visual classifiers in the presence of spurious correlations. We are inspired by prior work on machine learning fairness [51, 52, 41, 1] and aim to build a unified understanding of the proposed techniques. Code is available at https://github.com/princetonvisualai/DomainBiasMitigation.

We begin by proposing a simple but surprisingly effective benchmark for studying the effect of data bias on visual recognition tasks. Classical literature on mitigating bias generally operates on simpler (often linear) models [11, 50, 28], which are easier to understand and control; only recently have researchers begun looking at mitigating bias in end-to-end trained deep learning models [16, 2, 40, 18, 48, 25, 30, 36, 47, 17]. Our work helps bridge the gap, proposing an avenue for exploring mitigating bias in Convolutional Neural Network (CNN) models within a simpler and easier-to-analyze setting than with a fully-fledged black-box system. By utilizing dataset augmentation to introduce controlled biases, we provide simple and precise targets for model evaluation (Sec. 3).

Using this benchmark, we demonstrate that the presence

[1] Buolamwini and Gebru [6] note that collecting a more representative training dataset should be the first step of the solution. That is true in the cases they consider (where people with darker skin tones are dramatically and unreasonably undersampled in datasets) but may not be a viable approach to cases where the datasets accurately reflect the real-world skew.

of spurious bias in the training data severely degrades the accuracy of current models, even when the biased dataset contains strictly more information than an unbiased dataset. We then provide a thorough comparison of existing methods for bias mitigation, including domain adversarial training [46, 41, 1], Reducing Bias Amplification [52], and domain conditional training similar to [40]. To the best of our knowledge, no such comparison exists currently as these methods have been evaluated on different benchmarks under varying conditions and have not been compared directly. We conclude that a domain-independent approach inspired by [11] outperforms more complex competitors (Sec. 4).

Finally, we validate our findings in more realistic settings. We evaluate on the CelebA [32] benchmark for attribute recognition in the presence of gender bias (Sec. 5). We demonstrate that our domain-independent training model successfully mitigates real-world gender bias.

2. Related Work

Mitigating Spurious Correlation. Recent work on the effects of human bias on machine learning models investigates two challenging problems: identifying and quantifying bias in datasets, and mitigating its harmful effects. In relation to the former, [5, 31] study the effect of class-imbalance on learning, while [52] reveal the surprising phenomenon of bias amplification. Additionally, recent works have shown that ML models possess bias towards legally protected classes [29, 6, 4, 7, 33, 8]. Our work complements these by presenting a dataset that allows us to isolate and control bias precisely, alleviating the usual difficulties of quantifying bias.

On the bias mitigation side, early works investigate techniques for simpler linear models [23, 50]. Our constructed dataset allows us to isolate bias while not simplifying our architecture. More recently, works have begun looking at more sophisticated models. For example, [52] propose an inference update scheme to match a target distribution, which can remove bias. [40] introduce InclusiveFaceNet for improved attribute detection across gender and race subgroups; our discriminative architecture is inspired by this work. Conversely, [12] propose a scheme for decoupling classifiers, which we use to create our domain independent architecture. The last relevant approach to bias mitigation for us is adversarial mitigation [1, 51, 13, 16]. Our work uses our novel dataset to explicitly highlight the drawbacks, and offers a comparison between these mitigation strategies that would be impossible without access to a bias-controlled environment.

Fairness Criterion. Pinning down an exact and generally applicable notion of fairness is an inherently difficult and important task. Various fairness criteria have been introduced and analyzed, including demographic parity [24, 51], predictive parity [15], error-rate balance [19], equality-of-odds and equality-of-opportunity [19], and fairness-through-unawareness [35] to try to quantify bias. Recent work has shown that such criteria must be selected carefully; [19] prove minimizing error disparity across populations, even under relaxed assumptions, is equivalent to randomized predictions; [19] introduce and explain the limitations of an 'oblivious' discrimination criterion through a non-identifiability result; [35] demonstrate that ignoring protected attributes is ineffective due to redundant encoding; [11] show that demographic parity does not ensure fairness. We define our tasks such that test accuracy directly represents model bias.

Surveying Evaluations. We are inspired by previous work which aggregate ideas, methods and findings to provide a unify survey of a subfield of computer vision [22, 38, 43, 21]. For example, [45] surveys relative dataset biases present in computer vision datasets, including selection bias (datasets favoring certain types of images), capture bias (photographers take similar photos), category bias (inconsistent or imprecise category definitions), and negative set bias (unrepresentative or unbalanced negative instances). We continue this line of work for bias mitigation methods for modern visual recognition systems, introducing a benchmark for evaluation which isolates bias, and showing that our analysis generalizes to other, more complex, biased datasets.

3. A Simple Setting for Studying Bias

We begin by constructing a novel benchmark for studying bias mitigation in visual recognition models. This setting makes it possible to demonstrate that the presence of spurious correlations in training data severely degrades the performance of current models, even if learning such spurious correlations is sub-optimal for the target task.

CIFAR-10S Setup. To do so, we design a benchmark that erroneously correlates target classification decisions (what object category is depicted in the image) with an auxiliary attribute (whether the image is color or grayscale).

We introduce CIFAR-10 Skewed (CIFAR-10S), based on CIFAR-10 [27], a dataset with 50,000 32×32 images evenly distributed between 10 object classes. In CIFAR-10S, each of the 10 original classes is subdivided into two new domain subclasses, corresponding to color and grayscale domains within that class. Per class, the 5,000 training images are split 95% to 5% between the two domains; five classes are 95% color and five classes are 95% grayscale. The total number of images allocated to each domain is thus balanced. For testing, we create two copies of the standard CIFAR-10 test set: one in color (COLOR) and one in grayscale (GRAY). These two datasets are considered separately, and only the 10-way classification decision boundary is relevant.

Discussion. We point out upfront that the analogy between color/grayscale and gender domains here wears thin: (1) we consider the two color/grayscale domains as purely binary and disjoint whereas the concept of gender is more fluid; (2) a color/grayscale domain classifier is significantly sim-

pler to construct than a gender recognition model; (3) the transformation between color and grayscale images is linear whereas the manifestation of gender is much more complex.

Nevertheless, we adopt this simple framework to distill down the core algorithmic exploration before diving into the more complex setups in Sec. 5. This formulation has several compelling properties: (1) we can control the correlation synthetically by changing images from color to grayscale, maintaining control over the distribution, (2) we can guarantee that color images contain strictly more information than grayscale images, maintaining control over the discriminative cues in the images, and (3) unlike other datasets, there is no fairness/accuracy trade off since both are complementary. Furthermore, despite its simplicity, this setup still allows us to study the behavior of modern CNN architectures.

Key Issue. We ground the discussion by presenting one key result that is counter-intuitive and illustrates why this very simple setting is reflective of a much deeper problem. We train a standard ResNet-18 [20] architecture with a softmax and cross-entropy loss for 10-way object classification. Training on the skewed CIFAR-10S dataset and testing on COLOR images yields $89.0 \pm 0.5\%$ accuracy.[2] This may seem like a reasonable result until we examine that a model trained on an all-grayscale training set (so never having seen a single color image!) yields a significantly higher 93.0% accuracy when tested out-of-domain on COLOR images.

This disparity occurs because the model trained on CIFAR-10S learned to correlate the presence of color and the object classes. When faced with an all-color test set, it infers that it is likely that these images come from one of the five classes that were predominantly colored during training (Fig. 1). In a real world bias setting where the two domains correspond to gender and the classification targets correspond to activities, this may manifest itself as the model making overly confident predictions of activities traditionally associated with female roles on images of women [52].

4. Benchmarking Bias Mitigation Methods

Grounded with the task at hand (training recognition models in the presence of spurious correlations) we perform a thorough benchmark evaluation of bias mitigation methods. Many of these techniques have been proposed in the literature for this task; notable exceptions include prior shift inference for bias mitigation (Sec. 4.3), the distinction between discriminative and conditional training in this context (Sec. 4.4), and the different inference methods for conditional training from biased data (Sec. 4.4). Our findings are summarized in Table 1. In Sec. 5 we demonstrate how our findings on CIFAR10S generalize to real world settings.

Setup. To perform this analysis, we utilize the CIFAR-10S domain correlation benchmark of Sec. 3. We assume that

[2]We report the mean across 5 training runs (except for CelebA in Sec. 5.2). Error bars are 2 standard deviations (95% confidence interval).

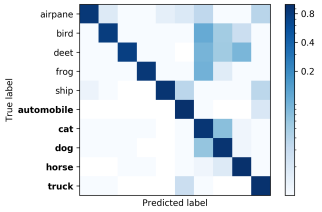

Figure 1. Confusion matrix of a ResNet-18 [20] classifier trained on the skewed CIFAR-10S dataset. The model has learned to correlate the presence of color with the five object classes (in bold) and predominantly predicts those classes on the all-color test set.

at training time the domain labels are available (e.g., we know which images are color and which are grayscale in CIFAR-10S, or which images correspond to pictures of men or women in the real-world setting). All experiments in this section build on the ResNet-18 [20] architecture trained on the CIFAR-10S dataset, with $N = 10$ object classes and $D = \{\text{color}, \text{grayscale}\}$. The models are trained from scratch on the target data, removing any potential effects from pretraining. Unless otherwise noted the models are tarined for 200 epochs, with SGD at a learning rate of 10^{-1} with a factor of 10 drop-off every 50 epochs, a weight decay of $5\text{e}{-4}$, and a momentum of 0.9. During training, the image is padded with 4 pixels on each side and then a 32×32 crop is randomly sampled from the image or its horizontal flip.

Evaluation. We consider two metrics: mean per-class per-domain accuracy (primary) and bias amplification of [52]. The test set is fully balanced across domains, so mean accuracy directly correlates with the model's ability to avoid learning the domain correlation during training. We include the mean bias metric for completeness with the literature, as

$$\frac{1}{|C|} \sum_{c \in C} \frac{\max(\text{Gr}_c, \text{Col}_c)}{\text{Gr}_c + \text{Col}_c} - 0.5. \tag{1}$$

where Gr_c is the number of grayscale test set examples predicted to be of class c, while Col_c is the same for color.

4.1. Strategic Sampling

The simplest approach is to strategically sample with replacement to make the training data 'look' balanced with respect to the class-domain frequencies. That is, we sample rare examples more often during training, or, equivalently, utilize non-uniform misclassification cost [14, 3]. However, as detailed in [49], there are significant drawbacks to oversampling: (1) seeing exact copies of the same example during training makes overfitting likely, (2) oversampling increases the number of training examples without increasing the amount of information, which increases learning time.

Experimental Evaluation. The baseline model first presented in Sec. 3 is a ResNet-18 CNN with a softmax clas-

Model Name	Model	Test Inference	Bias (↓)	Accuracy (%, ↑) Color	Gray	Mean
Baseline	N-way softmax	$\arg\max_y \mathrm{P}(y\|x)$	0.074	89.0	88.0	88.5 ± 0.3
Oversampling	N-way softmax, resampled	$\arg\max_y \mathrm{P}(y\|x)$	0.066	89.2	89.1	89.1 ± 0.4
Adversarial	w/ uniform confusion [1, 46]	$\arg\max_y \mathrm{P}(y\|x)$	0.101	83.8	83.9	83.8 ± 1.1
	w/ ∇ reversal, proj. [51]	$\arg\max_y \mathrm{P}(y\|x)$	0.094	84.6	83.5	84.1 ± 1.0
DomainDiscrim	joint ND-way softmax	$\arg\max_y \sum_d \mathrm{P}_{\mathrm{tr}}(y, d\|x)$	0.844	88.3	86.4	87.3 ± 0.3
		$\arg\max_y \max_d \mathrm{P}_{\mathrm{te}}(y, d\|x)$	0.040	91.3	89.3	90.3 ± 0.5
		$\arg\max_y \sum_d \mathrm{P}_{\mathrm{te}}(y, d\|x)$	0.040	91.2	89.4	90.3 ± 0.5
	RBA [52]	$y = \mathcal{L}(\sum_d \mathrm{P}_{\mathrm{tr}}(y, d\|x))$	0.054	89.2	88.0	88.6 ± 0.4
DomainIndepend	N-way classifier per domain	$\arg\max_y \mathrm{P}_{\mathrm{te}}(y\|d^*, x)$	0.069	89.2	88.7	88.9 ± 0.4
		$\arg\max_y \sum_d s(y, d, x)$	**0.004**	**92.4**	**91.7**	$\mathbf{92.0 \pm 0.1}$

Table 1. Performance comparison of algorithms on CIFAR-10S. All architectures are based on ResNet-18 [20]. We investigate multiple bias mitigation strategies, and demonstrate that a domain-independent classifier outperforms all baselines on this benchmark.

sication layer, which achieves $88.5 \pm 0.3\%$ accuracy. The same model with oversampling improves to $89.1 \pm 0.4\%$ accuracy. Both models drive the training loss to zero. Note that data augmentation is critical for this result: without data augmentation the oversampling model achieves only $79.2 \pm 0.8\%$ accuracy, overfitting to the data.

4.2. Adversarial Training

Another approach to bias mitigation commonly suggested in the literature is *fairness through blindness*. That is, if a model does not look at, or specifically encode, information about a protected variable, then it cannot be biased. To this end, adversarial training is set up through the minimax objective: maximize the classifier's ability to predict the class, while minimizing the adversary's ability to predict the protected variable based on the underlying learned features.

This intuitive approach, however, has a major drawback. Suppose we aim to have equivalent feature representations across domains. Even if a particular protected attribute does not exist in the feature representation of a classifier, combinations of other attributes can be used as a proxy. This phenomenon is termed *redundant encoding* in the literature [19, 11]. For an illustrative example, consider a real-world task of a bank evaluating a loan application, irrespective of the applicant's gender. Suppose that the applicant's employment history lists 'nurse'. It can thus, by proxy, be inferred with high probability that the applicant is also a woman. However, employment history is crucial to the evaluation of a loan application, and thus the removal of this redundant encoding will degrade its ability to perform the evaluation.

Experimental Evaluation. We apply adversarial learning to de-bias the object classifier. We consider both the uniform confusion loss $-(1/|D|)\sum_d \log q_d$ of [1] (inspired by [46]), and the loss reversal $\sum_d \mathbb{1}[\widehat{d} = d] \log q_d$ with gradient projection of [51].[3] These methods achieve only 83.4% and 84.1% accuracy, respectively. As Fig. 2 visually demonstrates, although the adversarial classifier enforces domain confusion it additionally creates undesirable class confusion.

We run one additional experiment to validate the findings. We test whether models encode the domain (color/grayscale) information even when *not* exposed to a biased training distribution; if so, this would help explain why minimizing this adversarial objective would lead to a worse underlying feature representation and thus reduced classification accuracy. We take the feature representation of a 10-way classifier trained on *all color* images (so not exposed to color/grayscale skew) and train a linear SVM adversary on this feature representation to predict the color/grayscale domain of a new image. This yields an impressive 82% accuracy; since the ability to discriminate between the two domains emerges naturally even without biased training, it would make sense that requiring that the model not be able to distinguish between the two domains would harm its overall classification ability.

4.3. Domain Discriminative Training

The alternative to fairness through blindness is *fairness through awareness* [11] where the domain information is first explicitly encoded and then explicitly mitigated. The simplest approach is training a ND-way discriminative classifier where N is the number of target classes and D is the number of domains. The correlation between domains and classes can then be removed during inference in one of several ways.

[3] We apply the adversarial classifiers on the penultimate layer for [1, 46] model, and on the final classification layer for [51] as recommended by the authors. We experimented with other combinations of layers and losses, including applying the projection method of [51] onto the confusion loss of [1, 46], and achieved similar results. The models are trained for 500 epochs using Adam with learning rates 3e-4 and weight decay 1e-4. We hold out 10,000 images to tune the hyperparameters before retraining the network on the entire training set. To verify training efficacy, we train SVM domain classifiers on the learned features: the accuracy is 99.0% before and 78.2% after adversarial training, verifying training effectiveness.

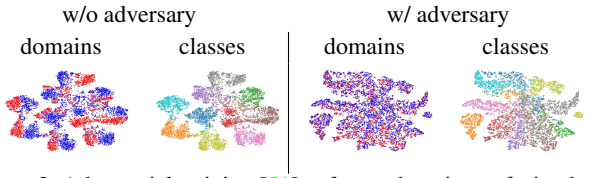

Figure 2. Adversarial training [51] enforces domain confusion but also introduces unwanted class boundary confusion (t-SNE plots).

4.3.1 Prior Shift Inference

If the outputs of the ND-way classifier can be interpreted as probabilities, a test-time domain solution to removing class-domain correlation was introduced in [42] and applied in [37] to visual recognition. Let the classifier output a joint probability $\mathrm{P}(y, d|x)$ for target class y, domain d and image x. We can assume that $\mathrm{P}_{\mathrm{tr}}(x|y, d) = \mathrm{P}_{\mathrm{te}}(x|y, d)$, i.e., the distribution of image appearance within a particular class and domain is the same between training and test time. However, $\mathrm{P}_{\mathrm{tr}}(d, y) \neq \mathrm{P}_{\mathrm{te}}(d, y)$, i.e., the correlation between target classes and domains may have changed. This suggests that the test-time probability $\mathrm{P}_{\mathrm{te}}(y, d|x)$ should be computed as:

$$\mathrm{P}_{\mathrm{te}}(y, d|x) \propto \mathrm{P}_{\mathrm{te}}(x|y, d)\mathrm{P}_{\mathrm{te}}(y, d) \quad (2)$$

$$= \mathrm{P}_{\mathrm{tr}}(x|y, d)\mathrm{P}_{\mathrm{te}}(y, d) \quad (3)$$

$$\propto \mathrm{P}_{\mathrm{tr}}(y, d|x)\frac{\mathrm{P}_{\mathrm{te}}(y, d)}{\mathrm{P}_{\mathrm{tr}}(y, d)} \quad (4)$$

In theory, this requires access to the test label distribution $\mathrm{P}_{\mathrm{te}}(y, d)$; however, assuming uncorrelated d and y at test time (unbiased $\mathrm{P}_{\mathrm{te}}(d|y)$) and mean per-class accuracy evaluation (uniform $\mathrm{P}_{\mathrm{te}}(y)$), $\mathrm{P}_{\mathrm{te}}(y, d) = \mathrm{P}_{\mathrm{te}}(d|y)\mathrm{P}_{\mathrm{te}}(y) \propto 1$.

Eqn. 4 then simplifies to $\mathrm{P}_{\mathrm{tr}}(y, d|x)/\mathrm{P}_{\mathrm{tr}}(y, d)$, removing the test distribution requirement. With this assumption, the target class predictions can be computed directly as

$$\hat{y} = \arg\max_y \max_d \mathrm{P}_{\mathrm{te}}(y, d|x) \quad (5)$$

or, using the Law of Total Probability,

$$\hat{y} = \arg\max_y \mathrm{P}_{\mathrm{te}}(y|x) = \arg\max_y \sum_d \mathrm{P}_{\mathrm{te}}(y, d|x). \quad (6)$$

Experimental Evaluation. We train a ND-way classifier (20-way softmax in our setting) to discriminate between (class, domain) pairs. This discriminative model with inference prior shift towards a uniform test distribution (Eqn. 4) followed by sum of outputs (Eqn. 6) achieves 90.3% accuracy, significantly outperforming the $88.5 \pm 0.3\%$ accuracy of the N-way softmax baseline. To quantify the effects of the two steps of inference: taking the highest output predictor rather than summing across domains (Eqn. 5) has no effect on accuracy because the two domains are easily distinguishable in this case; however, summing the outputs without first applying prior shift drops accuracy from 90.3% to 87.3%.

Finally, we verify that the increase in accuracy is not just the result of the increased number of parameters in the classifier layer. We train an ensemble of baseline models, averaging their softmax predictions: one baseline achieves 88.5% accuracy, two models achieve 89.6%, and only an ensemble of *five* baseline models (with 55.9M trainable parameters) achieve 90.0% accuracy on par with 90.3% accuracy of the discriminative model (with 11.2M parameters).

4.3.2 Reducing Bias Amplification

An alternative inference approach is Reducing Bias Amplification ("RBA") of Zhao et al. [52]. RBA uses corpus-level constraints to ensure inference predictions follow a particular distribution. They propose a Lagrangian relaxation iterative solver since the combinatorial optimization problem is challenging to solve exactly at large scale. This method effectively matches the desired inference distribution and reduces bias; however, the expensive optimization must be run on all test samples before a single inference is possible.

Experimental Evaluation. In the original setting of [52], training and test time biases are equal. However, RBA is flexible enough to optimize for any target distribution. On CIFAR-10S, we thus set the optimization target bias to 0 and the constraint epsilon to 5%. To make the optimization as effective as possible, we substitute in the known test-time domain (because it can be perfectly predicted) so that the optimization only updates the class predictions.

Applying RBA on the $\sum_d \mathrm{P}_{\mathrm{tr}}(y, d|x)$ scores results in 88.6% accuracy, a 1.3% improvement over the simpler $\arg\max_y \sum_d \mathrm{P}_{\mathrm{tr}}(y, d|x)$ inference but an insignificant improvement over 88.5% of the BASELINE model. Interestingly, we also observe that the benefits of RBA optimization are significantly lessened when prior shift is applied beforehand. For example, when using the $\sum_d \mathrm{P}_{\mathrm{te}}(y, d|x)$ post-prior shift scores, accuracy only improves negligibly from 90.3% using $\arg\max$ inference to 90.4% using RBA. Therefore, we conclude that applying RBA after prior shift is extraneous. However, the converse is not true as the best accuracy achieved by RBA without prior shift is significantly lower than the accuracy achieved with prior shift inference.

4.4. Domain Independent Training

One concern with the discriminative model is that it learns to distinguish between the ND class-domain case; in particular, it explicitly learns the boundary between the same class across different domains (e.g., cat in grayscale versus cat in color, or a woman programming versus a man programming). This may be wasteful, as the N-way class decision boundaries may in fact be similar across domains and the additional distinction between the same class in different domains may not be necessary. Furthermore, the model is necessarily penalized in cases where the domain prediction is challenging but the target class prediction is unambiguous.

This suggests training separate classifiers per domain. Doing this naively, however, as an ensemble, will yield poor performance as each model will only see a fraction of the data. We thus consider a shared feature representation with an ensemble of classifiers. This alleviates the data reduction problem for the representation though not for the classifiers.

Given the predictions $\mathrm{P}(y|d,x)$, multiple inference methods are possible. If the domain d^* is known at test time, $\hat{y} = \arg\max_y \mathrm{P}(y|d^*,x)$ is reasonable yet entirely ignores the learned class boundaries in the other domains $d \neq d^*$, and may suffer if some classes y were poorly represented within d^* during training. If a probabilistic interpretation is possible, then two inference methods are reasonable:

$$\hat{y} = \arg\max_y \max_d \mathrm{P}(y|d,x), \text{ or} \tag{7}$$

$$\hat{y} = \arg\max_y \sum_d \mathrm{P}(y|d,x)\mathrm{P}(d|x) \tag{8}$$

However, Eqn. 7 again ignores the learned class boundaries across domains, and Eqn. 8 requires inferring $\mathrm{P}(d|x)$ (which may either be trivial, as in CIFAR-10S, reducing to a single-domain model, or complicated to learn and implicitly encoding the correlations between y and d that we are trying to avoid). Further, in practice, while the probabilistic interpretation of a single model may be a reasonable approximation, the probabilistic outputs of the multiple independent models are frequently miscalibrated with respect to each other.

A natural option is to instead reason directly on class boundaries of the D domains, and perform inference as[4]

$$\hat{y} = \arg\max_y \sum_d \mathrm{s}(y,d,x), \tag{9}$$

where $s(y,d,x)$ are the network activations at the classifier layer. For linear classifiers with a shared feature representation this corresponds to averaging the class decision boundaries. We demonstrate that this technique works well in practice across both single and multi-label target classification tasks at removing class-domain correlations.

Experimental Evaluation. We train a model for performing object classification on the two domains independently. This is implemented as two 10-way independent softmax classifiers sharing the same underlying network. At training time we use knowledge of the image domain to only update one of the classifiers. At test time we apply prior shift to adjust the output probabilities of both classifiers towards a uniform distribution, and consider two inference methods. First, we use only the classifier corresponding to the test domain, yielding a low 88.9% accuracy as expected because it is not able to integrate information across the two domains (despite requiring specialized knowledge of the image domain). Instead, we combine the decision boundaries following Eqn. 9 and achieve 92.0% accuracy, significantly outperforming the baseline of $88.5 \pm 0.3\%$.

[4]Interestingly, under a softmax probabilistic model this inference corresponds to the geometric mean between $\{\mathrm{P}(y|d,x)\}_d$, which is a stable method for combining independent models with different output ranges.

4.5. Summary of Findings

So far we illustrated that the CIFAR-10S setup is an effective benchmark for studying bias mitigation, and provided a thorough evaluation of multiple techniques. We demonstrated the shortcomings of strategic resampling and of adversarial approaches for bias mitigation. We showed that the prior shift inference adjustment of output probabilities is a simpler, more efficient, and more effective alternative to the RBA technique [52]. Finally, we concluded that the domain-conditional model with explicit combination of per-domain class predictions significantly outperforms all other techniques. Table 1 lays out the findings.

Recall our original goal of Sec. 3 to train a model that mitigates the domain correlation bias in CIFAR-10S enough to classify color images of objects as well as a model trained on only grayscale images would. We have partially achieved that goal. The DOMAININDEPENDENT model trained on CIFAR-10S achieves 92.4% accuracy on color images, significantly better than $89.0 \pm 0.5\%$ of BASELINE and approaching $93.0 \pm 0.2\%$ of the model trained entirely on grayscale images. However, much still remains to be done. We would expect that a model trained on CIFAR-10S would take advantage of the available color cues and perform even better than 93.0%, ideally approaching 95.1% accuracy of a model trained on all color images. The correlation bias is a much deeper problem for visual classifiers and much more difficult to mitigate than it appears at first glance.

5. Real World Experiments

While CIFAR-10S proves to be a useful landscape for bias isolation studies, there remains the implicit assumption throughout that such findings will generalize to other settings. Indeed, it is possible that they may not due to the synthetic nature of the proposed bias generation. We thus investigate our findings in three alternative scenarios. First, in Sec. 5.1 we consider two modifications to CIFAR-10S: varying the level of skew beyond the 95%-5% studied in Sec. 4, and replacing the color/grayscale domains with more realistic non-linear transformations. After verifying all our findings still hold, in Sec. 5.2 we consider face attribute recognition on the CelebA dataset [32] where the presence of attributes, e.g., "smiling" is correlated with gender.

5.1. CIFAR Extensions

There are two key distinctions between the CIFAR-10S dataset studied in Sec. 4 and the real world scenarios where gender or race are correlated with the target outputs.

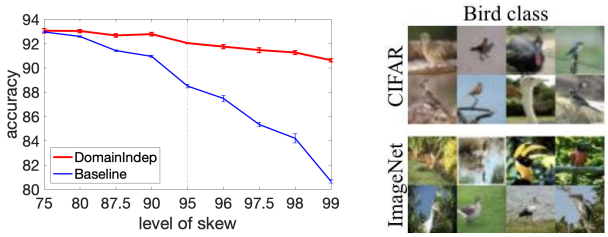

Figure 3. *(Left)* The DOMAINDEP model outperforms the BASELINE on CIFAR-10S for varying levels of skew. *(Right)* To investigate more real-world domains instead of color-grayscale, we consider the subtle shift between CIFAR and 32x32 ImageNet [10, 39].

MODEL	28x28crop	1/2 res.	1/4 res.	ImageNet
BASELINE	89.2	85.6	73.7	79.4
OVERSAMP	90.1	85.4	72.7	78.6
DOMDISCR	91.6	88.5	77.3	81.5
DOMINDEP	**93.0**	**90.2**	**79.9**	**83.5**

Table 2. On CIFAR-10S, we consider other transformations instead of the grayscale domain: (1) cropping the center of the image, (2,3) reducing the image resolution [44], followed by upsampling or (4) replacing with 32x32 ImageNet images of the same class [10]. We use the inference of Eqn. 6 for DOMDISCR and Eqn. 9 for DOMINDEP, and report mean per-class per-domain accuracy (in %). Our conclusions from Sec. 4 hold across all domain shifts.

Varying Degrees of Domain Distribution. The first distinction is in the *level* of skew, where domain balance may be more subtle than the 95%-5% breakdown studied above. To simulate this setting, we validated on CIFAR with different levels of color/grayscale skew, using the setup of Sec. 4 in Fig. 3 (*Left*). The DOMAINDEP model consistently outperforms the BASELINE, although the effect is significantly more pronounced at higher skew levels. For reference, the average gender skew on the CelebA dataset [32] for face attribute recognition described in Sec. 5.2 is 80.0%[5].

Other Non-Linear Transformations. The second distinction is that real-world protected attributes differ from each other in more than just a linear color-grayscale transformation (e.g., men and women performing the same task look significantly more different than the same image in color or grayscale). To approximate this in a simple setting, we followed the CIFAR protocol of Sec. 4, but instead of converting images to grayscale, we consider alternative domain options in Table 2. Arguably the most interesting shift corresponds to taking images of similar classes from ImageNet [39, 10], and we focus our discussion on that one.

The domain shift here is subtle (shown in Fig. 3 *Right*) but the conclusions hold: mean per-class per-domain accuracy is BASELINE $79.4 \pm 0.4\%$, ADVERSARIAL $74.1 \pm 0.6\%$ [1, 46] and $73.1 \pm 3.0\%$ [51] (not shown in Table 2), DOMAINDISCRIMINATIVE $81.5 \pm 0.7\%$, and our DOMAININDEPENDENT model $83.5 \pm 0.3\%$. One interesting change is that OVERSAMPLING yields $78.6 \pm 0.4\%$, significantly lower than the baseline of 79.4%, so we investigate further. The drop can be explained by the five classes which were heavily skewed towards CIFAR images at training time: the model overfit to the small handful of ImageNet images which got oversampled, highlighting the concerns with oversampling particularly in situations where the two domains are different from each other and the level of imbalance is high. We observe similar results in the high-to-low-resolution domain shift (third and fourth columns of Table 2), where the two domains are again very different from each other. To counteract this effect we instead applied the class-balanced loss method Cui et al. [9], cross-validating the hyperparameter on a validation set to $\beta = 0.9$, and achieved a more reasonable result of 79.2%, on par with $79.4 \pm 0.4\%$ of BASELINE but still behind $83.5 \pm 0.3\%$ of DOMAININDEPENDENT.

[5] In this multi-label setting the gender skew is computed on the dev set as the mean across 39 attributes of $\frac{\min(|attr=1,woman|,|attr=1,man|)}{|attr=1|}$.

5.2. CelebA Attribute Recognition

Finally, we verified our findings on the real-world CelebA dataset [32], used in [41] to study face attribute recognition when the presence of attributes, e.g., "smiling," is correlated with gender. We trained models to recognize the 39 attributes (all except the "Male" attribute). Out of the 39 attributes, 21 occur more frequently with women and 18 with men, with an average gender skew of 80.0% when an attribute is present. During evaluation we consider the 34 attributes that have sufficient validation and test images.[6]

Task and Metric. The target task is multi-label classification, evaluated using mean average precision (mAP) across attributes. We remove the gender bias in the test set by using a weighted mAP metric: for an attribute that appears with N_m men and N_w women images, we weight every positive man image by $(N_m + N_w)/(2N_m)$ and every positive woman image by $(N_m + N_w)/(2N_w)$ when computing the true positive predictions. This simulates the setting where the total weight of positive examples within the class remains constant but is now equally distributed between the genders.

We also evaluate the bias amplification (BA) of each attribute [52]. For an attribute that appears more frequently with women, this is $P_w/(P_m+P_w) - N_w/(N_m+N_w)$ where P_w, P_m are the number of women and men images respectively classified as positive for this attribute. For attributes that appear more frequently with men, the numerators are P_m and N_m. To determine the binary classifier decision we compute a score threshold for each attribute which maximizes the classifier's F-score on the validation set. Since our methods aim to de-correlate gender with the attribute we expect that bias amplification will be *negative* as the

[6] The removed attributes did not contain at least 1 positive male, positive female, negative male, and negative female image. They are: 5 o'clock shadow, goatee, mustache, sideburns and wearing necktie.

MODEL	MODEL	MAP	BA
BASE	N sigmoids	74.7	0.010
ADVER	w/uniform conf. [1, 46]	71.9	0.019
DOMDIS	2N sigm, $\sum_d \mathrm{P}_{\mathrm{tr}}(y, d\|x)$	73.8	0.007
DOMIND	2N sigmoids, $\mathrm{P}_{\mathrm{tr}}(y\|d^*, x)$	73.8	0.009
	2N sigm, $\max_d \mathrm{P}_{\mathrm{tr}}(y\|d, x)$	75.4	**-0.039**
	2N sigm, $\sum_d \mathrm{P}_{\mathrm{tr}}(y\|d, x)$	76.0	-0.037
	2N sigmoids, $\sum_d s(y, d, x)$	**76.3**	-0.035

Table 3. Attribute classification accuracy evaluated using mAP (in %, ↑) weighted to ensure an equal distribution of men and women appearing with each attribute, and Bias Amplification (↓). Evaluation is on the CelebA test set, across 34 attributes that have sufficient validation data; details in Sec. 5.2.

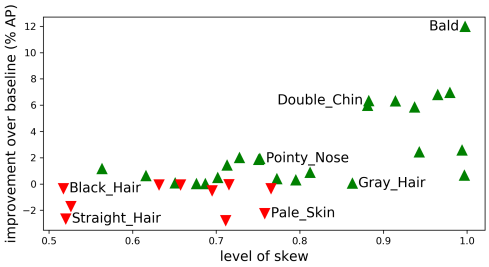

Figure 4. Per-attribute improvement of the DOMAININDEPENDENT model over the BASELINE model on the CelebA validation set, as a function of the level of gender imbalance in the attribute. Attributes with high skew (such as "bald") benefit most significantly.

predictions approach a uniform distribution across genders.

Training Setup. The images are the Aligned&Cropped subset of CelebA [32]. We use a ResNet-50 [20] base architecture pre-trained on ImageNet [39]. The FC layer of the ResNet model is replaced with two consecutive fully connected layers. Dropout and relu is applied to the output between the two fully connected layers, which has size 2048. It is trained with a binary cross entropy loss with logits using a batch size of 32, for 50 epochs with the Adam optimizer [26] (learning rate 1e-4). The best model over all epochs is selected per inference method on the validation set. For adversarial training, we run an extensive hyperparameter search over the relative weights of the losses and the number of epochs of the adversary. We select the model with the highest weighted mAP on the validation set among all models that successfully train a de-biased representation (accuracy of the gender classifier drops by at least 1%; otherwise it's essentially the BASELINE model with the same mAP). The models are evaluated on the test set.

Results. Table 3 summarizes the results. The overall conclusions from Sec. 4 hold despite the transition to the multi-label setting and to real-world gender bias. ADVERSARIAL training as before de-biases the representation but also harms the mAP (71.9% compared to 74.7% for BASELINE). In this multi-label setting we do not consider a probabilistic interpretation of the output as the classifier models are trained independently instead of jointly in a softmax. Without this interpretation and prior shift the DOMAINDISCRMINATIVE model achieves less competitive results than the baseline at 73.8%. RBA inference of [52] towards a uniform distribution performs similarly at 73.6%. The DOMAININDEPENDENT model successfully mitigates gender bias and outperforms the domain-unaware BASELINE on this task, increasing the weighted mAP from 74.7% to 76.3%. Alternative inference methods, such as selecting the known domain, computing the max output over the domains, or summing the outputs of the probabilities directly achieve similar bias amplification results but perform between $0.3 - 2.5\%$ mAP worse.

Analysis. We take a deeper look at the per-class results on the validation set to understand the factors that contribute to the improvement. Overall the DOMAININDEPENDENT model improves over BASELINE on 24 of the 34 attributes. Fig. 4 demonstrates that the level of gender skew in the attribute is highly correlated with the amount of improvement ($\rho = 0.709$). Attributes that have skew greater than 80% (out of the positive training images for this attribute at least 80% belong to one of the genders) always benefit from the DOMAININDEPENDENT model. This is consistent with the findings from CIFAR-10S in Fig. 3*(Left)*. When the level of skew is insufficiently high the harm from using fewer examples when training the DOMAININDEPENDENT model outweighs the benefit of decomposing the representation.

Oversampling. Finally, we note that the OVERSAMPLING model in this case achieves high mAP of 77.6% and bias amplification of -0.061, outperforming the other techniques. This is expected as we know from prior experiments in Sec. 4 and 5.1 that oversampling performs better in settings where the two domains are more similar (color/grayscale, 28x28 vs 32x32 crop) and where the skew is low while the dataset size is large so it wouldn't suffer from overfitting.

6. Conclusions

We provide a benchmark and a thorough analysis of bias mitigation techniques in visual recognition models. We draw several important algorithmic conclusions, while also acknowledging that this work does not attempt to tackle many of the underlying ethical fairness questions. What happens if the domain (gender in this case) is non-discrete? What happens if the imbalanced domain distribution is not known at training time – for example, if the researchers failed to identify the undesired correlation with gender? What happens in downstream tasks where these models may be used to make prediction decisions? We leave these and many other questions to future work.

Acknowledgements. This work is partially supported by the National Science Foundation under Grant No. 1763642, by Google Cloud, and by the Princeton SEAS Yang Family Innovation award. Thank you to Arvind Narayanan and to members of Princeton's Fairness in AI reading group for great discussions.

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCV*, 2018. 1, 2, 4, 7, 8

[2] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 1

[3] Steffen Bickel, Michael Brckner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, Sep 2009. 3

[4] 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. In *NeurIPS*, 2016. 1, 2

[5] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv preprint arXiv:1710.05381*, Oct. 2017. 2

[6] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018. 1, 2

[7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 1, 2

[8] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *ICML*, 2019. 2

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 7

[10] Luke Nicholas Darlow, Elliot J. Crowley, Antreas Antoniou, and Amos J. Storkey. CINIC-10 is not imagenet or CIFAR-10. *CoRR*, abs/1810.03505, 2018. 7

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Innovations in Theoretical Computer Science Conference*, 2012. 1, 2, 4

[12] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018. 2

[13] Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. In *ICLR*, 2016. 2

[14] Charles Elkan. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001. 3

[15] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017. 2

[16] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, 2015. 1, 2

[17] Aditya Grover, Kristy Choi, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. *arXiv preprint arXiv:1910.12008*, 2019. 1

[18] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *NeurIPS*, 2019. 1

[19] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *NeurIPS*, 2016. 2, 4

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 3, 4, 8

[21] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 2

[22] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2017. 2

[23] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the Damage of Dataset Bias. In *ECCV*, 2012. 2

[24] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NeurIPS*, 2017. 2

[25] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, 2019. 1

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 8

[27] Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. 2009. 2

[28] Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. *ICLR*, 2018. 1

[29] Sam Levin. A beauty contest was judged by AI and the robots didnt like dark skin, September 2016. 2

[30] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *CVPR*, 2019. 1

[31] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(2), 2009. 2

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2, 6, 7, 8

[33] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018. 2

[34] Safiya Umoja Noble. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, February 2018. 1

[35] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *KDD*, 2008. 2

[36] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *CVPR*, 2019. 1

[37] Amelie Royer and Christoph H Lampert. Classifier adaptation at prediction time. In *CVPR*, 2015. 5

[38] Olga Russakovsky, Jia Deng, Zhiheng Huang, Alexander C. Berg, and Li Fei-Fei. Detecting avocados to zucchinis: What have we done, and where are we going? In *ICCV*, 2013. 2

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 7, 8

[40] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018. 1, 2

[41] Hee Jung Ryu, Margaret Mitchell, and Hartwig Adam. Improving Smiling Detection with Race and Gender Diversity. *arXiv preprint arXiv:1712.00193*, 2017. 1, 2, 7

[42] 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, 2002. 5

[43] Gunnar A. Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *ICCV*. IEEE, Oct 2017. 2

[44] Jong-Chyi Su and Subhransu Maji. Adapting models to signal degradation using distillation. In *BMVC*, 2017. 7

[45] Antonio Torralba and Alexei A. Efros. Unbiased Look at Dataset Bias. In *CVPR*. IEEE, 2011. 2

[46] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *CVPR*, 2015. 2, 4, 7, 8

[47] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *CVPR*, 2019. 1

[48] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *CVPR*, 2019. 1

[49] Gary M Weiss, Kate McCarthy, and Bibi Zabar. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? 3

[50] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *ICML*, 2013. 1, 2

[51] 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*, 2018. 1, 2, 4, 5, 7

[52] 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 *EMNLP*, 2017. 1, 2, 3, 4, 5, 6, 7, 8