Overwriting Pretrained Bias with Finetuning Data

Angelina Wang
Princeton University
angelina.wang@princeton.edu

Olga Russakovsky
Princeton University
olgarus@princeton.edu

Abstract

Transfer learning is beneficial by allowing the expressive features of models pretrained on large-scale datasets to be finetuned for the target task of smaller, more domain-specific datasets. However, there is a concern that these pretrained models may come with their own biases which would propagate into the finetuned model. In this work, we investigate bias when conceptualized as both spurious correlations between the target task and a sensitive attribute as well as underrepresentation of a particular group in the dataset. Under both notions of bias, we find that (1) models finetuned on top of pretrained models can indeed inherit their biases, but (2) this bias can be corrected for through relatively minor interventions to the finetuning dataset, and often with a negligible impact to performance. Our findings imply that careful curation of the finetuning dataset is important for reducing biases on a downstream task, and doing so can even compensate for bias in the pretrained model.

1. Introduction

The current paradigm in machine learning typically involves using an off-the-shelf pretrained model that has been trained on a large-scale dataset, and then finetuning it on a smaller, application-specific dataset. This transfer learning is especially common for high-dimensional data like images and language [14, 11, 62]. However, large-scale datasets have been criticized for their biases [38, 59, 4, 9], which leaves open the concern that models pretrained on such datasets may carry biases over into the finetuned model. On the other hand, pretraining has been shown to confer benefits in model robustness and uncertainty estimation [24], so there is also the potential that pretrained models can reduce downstream biases by being more regularized or resistant to spurious correlations [5]. In our work, we investigate the implications of bias in pretrained models for the downstream finetuning task, and provide actionable insights on how to counteract this.

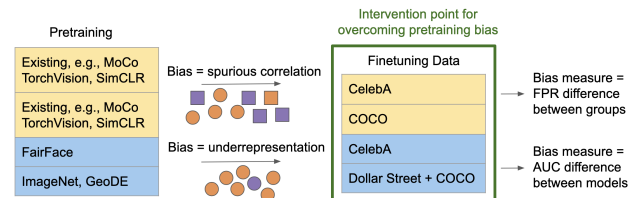We have been deliberately vague thus far on what we

Figure 1. We explore bias transference from pretrained to finetuned models in two forms in this work: spurious correlations and underrepresentation. We find that intervening on the finetuning data allows us to overcome bias from pretraining, often without compromising on performance.

mean by "bias." In this work, we operationalize bias in two ways based on what has been found thus far to be problematic in image features: *spurious correlations* between a sensitive attribute and target task [65, 60, 51, 61] and reduced performance from *underrepresentation* [6, 10, 50].

The topic of whether pretrained biases matter in finetuning is often assumed to be obvious, with contradictory arguments containing intuitively plausible explanations on both sides of the debate: that it does matter because the pretrained model brings biased features [44], or it does not because finetuning data will overwrite any pretrained biases [7, 15, 56]. Due to this uncertainty, it is not clear how to react to biases found in the features of pretrained models [55, 52, 17]. Of course, there is not a singular binary answer, as much is dependent upon the particulars of the training task. However, we bring much-needed clarity to the space for computer vision tasks, and give advice about bias transference from using pretrained models.

In this work we study the two notions of bias, *spurious correlations* and *underrepresentation*, by finetuning a variety of different pretrained models (Fig. 1). For each notion of bias, we first show our results on the CelebA dataset [32]. Then for spurious correlations we investigate the more complex COCO dataset [31] using real-world popular pretrained models (e.g., **MoCo** [21], **SimCLR** [8]). For underrepresentation, we look to the Dollar Street dataset [41] using pretrained models of our own design to test for specific hypotheses. On both forms of bias we find the following: 1) models finetuned on top of pretrained models can inherit

their biases (for spurious correlations, this is especially true if the correlation level is high, the salience of the bias signal is high relative to the true task signal, and/or the number of finetuning samples is low); 2) this bias can be relatively easily corrected for by curating the distribution of the finetuning dataset, *with a negligible impact to performance*. For example on CelebA, we find that by manipulating the strength of the spurious correlation in the finetuning dataset from 20% to 30%, we can retain the same high performance from using a biased pretrained model, but cut the amount of bias almost in half. The implications for this are significant: practitioners can use the pretrained model that lends the best performance in most cases so long as they appropriately curate the finetuning dataset, and thus get the best of both worlds in terms of performance and fairness. This means that significant consideration and effort needs to be spent on the curation of finetuning datasets, in a way that may not necessarily reflect the distribution of the test set, in order to "correct" for the biases of the pretrained model.[1]

2. Related Work

Rather than exploring all viable algorithmic bias mitigation strategies, we focus in this work on studying the already sufficiently-complex effect of the finetuning dataset's composition. As datasets grow increasingly larger in size and harder to control, we position our work as directing attention to the finetuning dataset (as opposed to the pretraining dataset) as the better site for investigation and mitigation.

Benefits of Pretraining. Pretrained features can be learned through methods like unsupervised [8, 21], self-supervised [35, 12], or supervised learning [36]. While there are indisputable time- and compute-saving benefits to transfer learning, He et al. [22] showed that sometimes models trained from scratch can match the performance of models finetuned from pretrained weights. However, even in those cases, Hendrycks et al. [24] finds that the finetuned model has superior robustness and uncertainty estimates. In our work, we try to understand whether pretrained models confer any other such benefits or harms in terms of bias.

Identifying Pretraining Bias. Prior works have sought to measure the fairness of image features [17, 55, 52], the concern being that any biases in image features may propagate into predictions based on such features. Goyal et al. [16] report that pretrained features trained without supervision are more robust and fair than those with supervision, and Sirotkin et al. [52] find that within self-supervised models, contrastive losses lead to more biases—however, neither of these works measure bias on a downstream task.[2] So, while

the pretrained features may be biased, it still does not tell us the implications for a downstream task.

Prior work in NLP has measured the correlation of "intrinsic" metrics like biases in the word embedding to "extrinsic" metrics like biases on the downstream task which uses these word embeddings, finding little correlation between these metrics [15, 7]. Other work in NLP has found that the distribution of the finetuning dataset matters more than the pretraining dataset in terms of bias [56, 53]. We find similar results in our study on pretrained image models, and give advice on how to correct for this. There is also work beginning to study the multimodal pretraining biases in the vision language domain [54].

Bias Transferrence in Vision. Salman et al. [44] similarly study bias transference, but focus on an operationalization of bias akin to backdoor attacks [18]. In their work and ours, we find that biases in pretrained models can propagate to finetuned models, but in our work we take this a step further by productively demonstrating that relatively minor interventions on the finetuning dataset can counteract this. While they find that in certain settings the biases in pretrained models persist even when the downstream dataset does not contain such biases, this different conclusion is likely due to their freezing of model layers. Kirichenko et al. [26] find that re-training just the last layer of a network can help models overcome spurious correlations. This matches the setting of one of our two conceptualizations of bias, and we add new findings when bias is underrepresentation. While their claim is stronger because they only re-train the last layer whereas we retrain the entire network, the setting we consider is very different because in our scenarios the available number of finetuning samples is relatively small, e.g., 128 or 1024 on CelebA, whereas in their comparable experiments they have access to the entire CelebA dataset (130, 216 samples). We find our setting of smaller finetuning datasets to be more realistic to many real-world scenarios.

3. Preliminaries

In this work, we will be conceptualizing "bias" in its two most studied forms in computer vision: *spurious correlations* and *underrepresentation*. Each of our experimental setups will consist first of a *pretraining* stage on the upstream task and then a *finetuning* stage on the downstream task. The *pretraining* stage can include simply instantiating a randomly initialized model (**Scratch**), instantiating the model weights to be that of an existing pretrained model (e.g., **SimCLR** [8], **MoCo** [21]), or pretraining our own model (e.g., **Gendered**). All pretrained models will be denoted in bold text. In the *finetuning* stage we train all layers of our model on the downstream task. We do not experi-

---

[1]Code at `https://github.com/princetonvisualai/overcoming-pretraining-bias`.

[2]While Goyal et al. [16] finetune their pretrained models on ImageNet for certain measurements, this is only to be able to extract labels from a

self-supervised model and ultimately assess those features.

ment with any freezing of layers, and thus our results serve as the upper bound for the effect that finetuning can have.

3.1. Datasets

We use six datasets in this work. They serve as training data for our upstream pretrained model, training data for our downstream finetuned model, or both. We use gender as our sensitive attribute for our analysis on spurious correlations, and use the terms `Men` and `Women` when referring to the annotated gender groups.[3] Our datasets are:

- CelebA [32]: 39 attributes such as `Brown Hair` and `High Cheekbones`; we use the label of `Male` for our `Men` group and not-`Male` for our `Women` group.

- FairFace [28]: racial attributes such as `White`; we use the labels of `Male` for `Men` and `Female` for `Women`.

- COCO [31]: 80 objects such as `Toothbrush` and `Oven`; we follow from prior work [65, 64] and derive gender labels based on the presence of gendered words in the captions. Even though COCO is most often a multi-label task, in some experiments we perform binary prediction on one label at a time to better isolate and manipulate the correlation of that particular object. Most images are from the Global North [10].

- ImageNet [42]: 200 objects such as `Fence` and `Barn`; most images are from the Global North [10].

- DollarStreet [41, 10]: 135 objects such as `Books` and `Motorcycles`, of which 15 map onto COCO object labels; this dataset is more geographically diverse than, e.g., COCO and ImageNet.

- GeoDE [40]: 40 objects such as `Bicycle` and `Jug`; deliberately collected to be geographically diverse.

A summary of how each dataset is used is in Fig. 1.

3.2. Implementation Details

We use a ResNet50 [23] with SGD for all of our experiments. The images are resized to be 224 by 224, normalized to ImageNet's mean pixel values, and randomly flipped horizontally for data augmentation. The particular hyperparameters chosen for finetuning are impactful [29], so we perform grid search for the lowest validation loss across the following hyperparameters: learning rate in $\{.1, .05, .01, .005, .001\}$ and weight decay in $\{0., .0001, .0005, .001\}$.[4] We train five random runs of every model to generate 95% confidence intervals. When we

need discrete predictions, i.e., $\{0, 1\}$, as opposed to continuous ones, i.e., $[0, 1]$, we pick the classification threshold for each label to be well-calibrated such that the percentage of predicted positive labels is the same as the true percentage.

4. Bias as Spurious Correlation

We begin our analysis by considering bias in the form of spurious correlations between the target label and a sensitive attribute which is predictive on the training set but not necessarily so on the test set [65, 60, 51, 61, 43]. We first demonstrate that finetuned models can inherit bias of this form from pretrained models on the CelebA dataset (Sec. 4.1). In other words, when the presence of a target label (e.g., `Eyebags`) may be erroneously learned to be correlated with gender, such that the classifier over-predicts `Eyebags` (i.e., a false positive) on images with `Women`. We thus measure the bias of spurious correlations as false positive rate (FPR) difference between `Women` and `Men`.[5] We show that pretrained bias is especially likely to affect the downstream task when the finetuning dataset has a high correlation with the sensitive attribute, low salience relative to the sensitive attribute, and/or a low number of finetuning samples.

Then we examine bias mitigation through finetuning on different distributions of the downstream dataset (Sec. 4.2). We show on two domains (CelebA attribute classification [32] and COCO object recognition [31]) that interventions on the finetuning dataset can mitigate much of the bias while retaining the performance gains of a biased pretrained model.

4.1. Finetuned models inherit spurious correlations from pretrained models

Initial Look. To start off, we want a set of pretrained models of differing biases to finetune on a set of downstream tasks where we can measure bias (i.e., FPR difference between `Women` and `Men`) in order to assess whether it makes a difference which pretrained model was used. For our downstream tasks, we perform binary classification on 11 different CelebA attributes.[6] To understand whether the bias of a pretrained model (i.e., learned correlation between gender and target attribute) will affect the bias on a finetuning task (i.e., applied correlation of gender and target at-

---

[3]Much fairness work is limited by the availability of sensitive attribute annotations [1], and our annotations generally treat gender as binary, a schema which harmfully alienates and erases different communities [47].

[4]For similar settings, e.g., same task with a different dataset distribution, we reuse the same hyperparameters in order to save computation time.

[5]Though often called equal opportunity, we will stay away from this term to not detract from its namesake [20], which is focused on allocational harms, when our application likely leads to representational harms [3].

[6]We selected the 11 attributes as follows: taking gender as our sensitive attribute, we are left with 39 in CelebA. Following from Ramaswamy et al. [39], we winnow this down to the 26 that do not suffer from severe data imbalance (i.e., positive label rate between 5-95%), then the 20 that are consistently labeled (e.g., `Black Hair` and not `Big Nose`), then the 14 that look the same between people of different genders (e.g., `Eyeglasses` and not `Young`). From here, we use three as to train our **Control** models, and are left with 11 to use as downstream tasks.
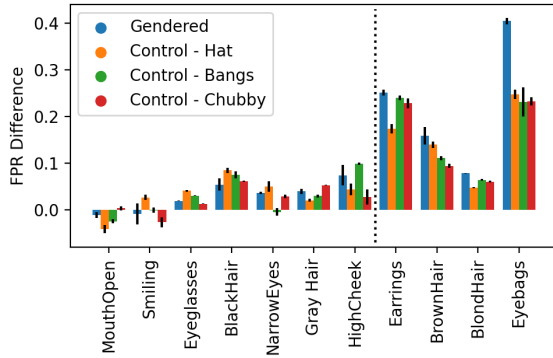
Figure 2. Across 11 CelebA attributes, the FPR difference between Women and Men of each model finetuned on a different base. The attributes are sorted by the ratio of the FPR difference between the **Gendered** model and highest **Control** model. The dotted line indicates the cutoff between the seven attributes that do not have a higher FPR difference from biased pretraining and the four attributes that do.

tribute during inference), we compare a pretrained model of our own design with high bias, called **Gendered**, to three of lesser bias, called **Control**. We then finetune all pretrained models and compare their downstream FPR gender difference. We create **Gendered** by training a model from scratch to classify gender (Men or Women) on the CelebA dataset.[7] As our **Control** pretrained models, we create three "less biased" ones that have been trained to classify the attributes of Wearing Hat, Bangs, and Chubby. These have been sparsely sampled to be of different saliences (how discriminable an attribute is).[8]

To understand whether finetuning on the **Gendered** model will cause more downstream bias than on a **Control** model, we compare the ratio of the FPR difference between the finetuned **Gendered** model and the most biased of the three finetuned **Control** models. As transfer learning is often done when there is insufficient data in the downstream domain, we do this on our 11 attributes using 1024 finetuning samples. In Fig. 2 we see that for 7 of our 11 attributes, finetuning on the **Gendered** model is no more biased than finetuning on any of the **Control** models; however for 4 of the attributes (Earrings, Brown Hair, Blond Hair, Eyebags), the pretrained model impacts downstream bias.

To understand what it is that differentiates the scenarios where pretrained bias does matter (4 of the 11 attributes) and where it does not (7 of the 11 attributes), we consider the relevance of three factors: (1) correlation level, (2) salience, and (3) number of finetuning samples. We show

CelebA results, with COCO results in the Supplementary.

(1) Correlation level and (2) Salience. Correlation level is the strength of the correlation between gender and the target attribute, and captures how useful it is for the upstream task's spurious correlation to be retained for the downstream task; salience is like the notion of discriminability, and captures whether an attribute is easier to learn in the downstream task than the spurious gender signal from the upstream task.[9] For quantifying these two concepts we adopt the methods from prior work [65, 39]. We calculate correlation level by $\frac{N(\text{Women, attribute})}{N(\text{Women, attribute})+N(\text{Men, attribute})}$ [39] where $N(\text{Women}, \text{attribute})$ indicates the number of images which are labeled as Women and have the attribute. Salience is measured relative to gender in this dataset, and positive values indicate gender is more salient than the downstream attribute (e.g., Mouth Slightly Open is not that salient), whereas negative values indicate it is less so (e.g., Eyeglasses is very salient). This value is calculated by constructing a setting where gender and the attribute are exactly correlated, and the salience score is derived from which of the gender or the attribute is learned better, details in Supplementary.

For any given attribute of a particular salience, we can artificially manipulate its correlation level by selecting different subsets of the data. We do so for all attributes to be 80%, and perform a mixed-effect regression analysis on our 22 observations (all 11 attributes of naturally different saliences at both their natural correlation level and at 80%), with parameter estimations adjusted by the group random effects for each attribute.[10]

We find the effect size for correlation level to be 2.19 (95% CI [.53, 3.85]) and salience to be 1.30 (95% CI [.27, 2.34]), both with $p < .05$. The positive coefficient for correlation level indicates that the stronger the correlation between the target task and gender (i.e., the more it benefits a model to rely on a gendered correlation), the more it will be affected by a biased pretrained model. The positive coefficient for salience indicates that attributes of lower salience are more likely to be affected by a biased pretrained model, likely because the gender visual cue is easier to spuriously rely on compared to the true downstream task cue.

Having seen that both correlation level and salience are able to explain part of why bias is transferred, we next investigate the effect of the number of finetuning examples.

(3) Finetuning number. We analyze the area under the ROC curve (AUC) and FPR difference of three pretrained models (**Scratch**, **Gendered**, **Control**) when finetuned on increasing amounts of data: $[2^4, 2^5, 2^6, ..., 2^{12}, 2^{13}, \text{full} =$

---

[7]We do not condone the use of gender predictors [19]. These models merely represent our conception of the kinds of image features that would be present in a model which has severe gendered correlations.

[8]We use three **Control** models of different saliences because, as we will show later, salience is a relevant factor in the transference of bias, and we want to ensure robust comparisons to a reasonable control condition.

[9]The notion of salience is related to simplicity bias [49, 57, 2, 34] in that neural networks may be more inclined to learn certain predictive patterns due to their being more salient or simple than others.

[10]We normalize correlation level to be between [0, 1], and calculate the FPR difference sign depending on which group the correlation is with.

130k]. To pick the comparison **Control** from our three, we select the one with the highest AUC at 2^{10} for each attribute.
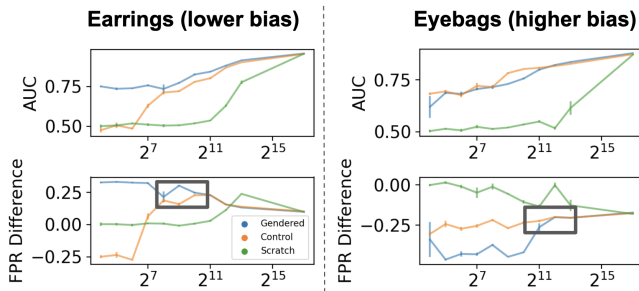

Figure 3. On two attributes from CelebA, we track the AUC and FPR difference between `Women` and `Men` for three pretrained bases (**Gendered**, **Control**, and **Scratch**) on increasing numbers of finetuning samples. The **Gendered** model typically has a higher magnitude FPR difference than the **Control** model, until the two converge with sufficient finetuning samples (gray box). `Earrings`, which we saw to be less susceptible to bias from a pretrained model in Fig. 2 and is of higher salience and less correlation level, achieves this at around 512 images, while `Eyebags` takes until around 4096.

In Fig. 3 we present results on two attributes which were more biased when finetuned on the **Gendered** model compared to the **Control**: `Earrings` which had the least bias and `Eyebags` which had the most. Up until the full finetuning dataset is used, the **Gendered** and **Control** models which have received pretraining are able to achieve better performance than the **Scratch** model, which is in turn fairer. While with sufficient finetuning samples all three models converge to the same levels of performance and bias, we may not always have a large number of finetuning samples. `Earrings`, which has higher salience and less correlation level than `Eyebags`, requires a smaller number of finetuning samples for it not to matter in terms of FPR difference whether the **Gendered** or **Control** pretrained model was used (around 512 compared to around 4096). When the fairnesses converge, it becomes negligible which pretrained model was used. This is a powerful result, as it indicates we can retain potential performance gains of a pretrained model, without needing to inherit its potential biases. While using a **Scratch** model would have alleviated concerns about the bias a pretrained model, it suffers too drastically in performance in the low data regime to always be feasible. In the rest of this work, we study how we might be able to retain the performance gains of different pretrained models while not inheriting their harmful biases.

4.2. Bias from spurious correlations can be corrected for in finetuning

We have now established that high correlation level, low salience, and low finetuning numbers are all relevant factors in terms of whether the bias of a pretrained model

will propagate into a finetuned model. With the exception of salience, which is hard to manually manipulate so we merely observe as a covariate, these other two factors become the levers we intervene on in our experiments going forward.

The scenario we consider is as follows: we have a dataset for a downstream task that is small such that we would like to leverage transfer learning from a pretrained model. However, we believe the pretrained model is likely to contain biases in the form of spurious correlations that could impact the downstream task. We want to understand whether there is a way we can benefit from the performance gains that the pretrained model brings, without inheriting all of the biases.

We investigate this on two datasets, CelebA and COCO, and use five pretrained models: **Scratch**, **TorchVision** [36], **MoCo** [21], **SimCLR** [8], and **Places** [66]. Each is finetuned on the downstream task, which has been artificially manipulated to have a correlation level of either 20% (towards men) or 80% (towards women), and the dataset is balanced to have equal numbers of images for the two genders. Without changing the distribution of this downstream test set, we artificially manipulate the downstream training set, i.e., finetuning dataset, across 11 increments of correlation level ([0%, 10%, ..., 100%]). We then assess whether there is a manipulated setting of the finetuning dataset such that a pretrained model is able to retain most of the performance gains it brings, while having less of a FPR difference than if trained directly on the natural finetuning distribution.

CelebA. We use two of the four attributes most susceptible to gender biases from the previous section (`Blond Hair` set to correlation level 80% and `Eyebags` set to correlation level 20%), and show the remaining two in the Supplementary. We visualize the tradeoff between performance (AUC) and bias (FPR difference) on the left of Fig. 4, and find both attributes have a configuration of the finetuning dataset which is different than the test dataset such that the performance is comparable to the peak performance, but the bias is significantly lower. For example, on `Eyebags`, **MoCo** has the highest performance but also the highest bias. However, by manipulating the finetuning correlation level from 20% to 30% the AUC remains at 0.88 while the FPR difference improves from -0.46 to -0.27. If we further manipulate the correlation level to 40%, the AUC drops slightly to 0.85, but the FPR difference improves to -0.05.

If we had restricted ourselves only to those models which are finetuned on a dataset that has the same correlation level as the test set (i.e., the bolded points), we would have had to compromise on performance in order to achieve lower bias. However, by manipulating the finetuning dataset such that it contains less spurious correlations than the test dataset, we can actually train a model that retains most of the performance benefits afforded to the more biased model, but also has the fairness of a less accurate model. This change
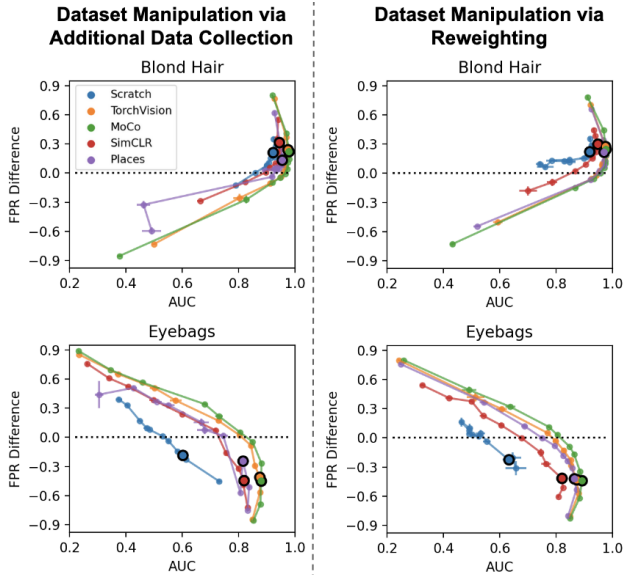
Figure 4. The performance and bias with 95% confidence intervals of pretrained models finetuned on different versions of two downstream tasks on CelebA: `Blond Hair` (correlated with women) and `Eyebags` (correlated with men). The bolded point indicates when the finetuning distribution matches the test distribution, and all other points indicate variations on the finetuning dataset. The left column represents when additional data is collected to manipulate the correlation level while maintaining 1024 finetuning samples, and the right column represents when finetuning data is reweighted such that the correlation level changes. On both the left and right, there are versions of the finetuning dataset that allow us to retain most performance gains while improving fairness.

is often relatively small as well, with just a 20% shift in the correlation. In the Supplementary we show results on 128 finetuning samples, where stronger manipulations are required to decrease the FPR difference.

Of course, these gains are not free: the added cost of this approach is the need to collect additional training samples to create the new correlation level. However, even in a constrained setting where you are unable to collect additional finetuning samples, reweighting the existing data to create an artificially less correlated dataset can bring about much of the same benefits. The right side of Fig. 4 represents this scenario where we are unable to collect additional samples, and simply create different correlation levels by reweighting samples. For `Eyebags`, finetuning **MoCo** by reweighting samples such that the correlation level goes from 20% to 30% barely drops the performance from 0.892 to 0.886, while the FPR difference improves from -0.45 to -0.34. If we make the larger change to 40% the performance drops to 0.86, but the FPR difference is improved to -0.17.

COCO. We investigate the two objects with the highest representation across both genders, because in this dataset the labels are more sparse and we want sufficient positive exam-
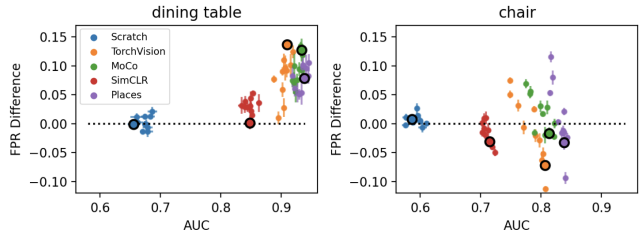

Figure 5. The performance and bias with 95% confidence intervals of pretrained models finetuned on different versions of two downstream tasks on COCO: `dining table` (correlated with women) and `chair` (correlated with men). The bolded point indicates when the finetuning distribution matches the test distribution, and all other points indicate variations on the finetuning dataset; all models are trained on 5000 finetuning samples. In each case, there is a version of the finetuning dataset that is different in distribution from the test set which allows us to train a model which retains the high AUC of a more biased model, but has lower FPR difference.

ples to be able to control the correlation level of an object with each gender. These two objects are `dining table`, which like the previous section we set to correlation level of 80%, and `chair`, which we set to 20%. We show results on additional objects in the Supplementary.

In Fig. 5 we show results on experiments with 5000 finetuning samples for the same setup as on CelebA.[11] Because of COCO's far noisier gender cues and additional complexity due to a greater diversity of image appearances, compared to CelebA's more uniform frontal facing individuals, the trends are not as clear where each change in correlation level directly corresponds to a reduction in FPR difference. However, what remains clear is there always exists a version of the finetuning dataset which retains the high AUC of a biased pretrained model, but reduces the bias on the downstream task. For example, on `chair` the finetuned **MoCo** has an AUC of 0.81 and FPR difference of -0.015; meanwhile, the finetuned **Places** has a better AUC of 0.84 but a worse FPR difference of -0.024. However, instead of having to choose between lower bias or higher performance, we find that by finetuning the better-performing **Places** model on a dataset of correlation level 40% rather than 20%, the performance stays constant at 0.84, while the FPR difference improves to -0.002! This indicates that careful curation of a finetuning dataset can overcome many FPR difference concerns, without compromising on the high performance that a pretrained model may be selected for.

In this section we have showed how bias in the form of spurious correlations can propagate from a pretrained model to a finetuned one. However, we have also shown that manipulations to the finetuning dataset, even if they cause the finetuning dataset to deviate in distribution from the downstream test set, can actually correct for much of

---

[11] We show results with 1000 finetuning examples in the Supplementary.

this bias while retaining high performance. While we have focused on the sensitive attribute of gender due to the availability of annotations, our findings are not necessarily restricted to this domain. Ultimately the spurious correlations we study in this section are simplified from the multi-attribute, multi-task ones that are likely to arise in the real world [61, 30, 63]. We did this to make the problem tractable to study with sufficient depth, and leave further exploration for future work. In the next section, we show similar results when bias is conceptualized as underrepresentation.

5. Bias as Underrepresentation

Now, we consider the implications of biased pretraining when "bias" means that one appearance of an object is underrepresented. This is inspired from both Buolamwini and Gebru [6], who showed that an underrepresentation of darker-skinned individuals led to worse classification performance for this group, and DeVries et al. [10], who showed that appearance differences, i.e., subcategories, within an object class led to objects from countries with lower household incomes to be misclassified more often. For example, bar soap is less recognizable as soap than pump soap is. In this section, we want to understand whether a pretrained model that has only learned about one subcategory of an object (e.g., pump soap), will perform worse on the other subcategory of an object (e.g., bar soap), compared to if it had seen that subcategory during pretraining. When we find that it does, we provide insight about the level of intervention on the finetuning dataset we can perform to overcome it.

5.1. Finetuned models do worse on subcategories underrepresented in pretrained models

We consider a downstream task `Target` to be composed of two possible subcategories: `T1`, e.g., pump soap, and `T2`, e.g., bar soap. We create **Pretrain-T1** that has only been trained to classify `T1` on FairFace and **Pretrain-T2** that has only been trained to classify `T2` on FairFace. Taking `T2` to be the underrepresented subcategory that we are particularly interested in the performance of, our measure of bias is the AUC difference on `T2` between **Pretrain-T2** and **Pretrain-T1**, i.e., the performance lost by having used a pretrained model which has not seen this subcategory before.[12]

For our downstream task, we use the CelebA dataset, and simulate the different appearances of an object through two different attributes as the subcategories of the classification label. For example, "`Light Hair`" could be composed

---

[12]AUC difference is between models rather than groups because different subcategories of an object may be inherently harder to classify or not, and we want to capture the relevant aspect of performance. Because FPR difference can be manipulated through post-hoc threshold changes [20], that comparison remains between groups.

of `Blond Hair` and `Brown Hair`. On 12 such subcategory pairings, selected to be roughly representative in terms of relative salience, we find that when we finetune on 128 images where half the positive labels are `T1` and half are `T2`, our bias measure of `T2` AUC difference is $.124 \pm .023$ between the two different pretrained models. The statistically significant positive difference indicates that a finetuned **Pretrain-T1** is not able to reach the performance on `T2` that a finetuned **Pretrain-T2** is. Experiment details and results on the analogous three factors of correlation level, salience, and finetuning number are in the Supplementary.

5.2. Bias from underrepresentation can be corrected for in finetuning

We have just established that a pretrained model (e.g., **Pretrain-T1**) which has not been trained on a particular label appearance (e.g., `T2`) will perform worse on it at finetuning time than if it had been trained on it (e.g., **Pretrain-T2**). In this section, we treat the proportion of finetuning dataset that is `T1` or `T2` as our analog to correlation level, and investigate the effect of manipulating it. It is obvious that increasing the proportion of `T2` will likely increase the performance on `T2`; however, it is not obvious what amount of proportion manipulation is required to compensate for a pretraining model not having seen a particular attribute.

Our experimental setup is as follows: the positive labels in our downstream task are 90% `T1` and 10% `T2`. As we know from the previous section, if we finetune **Pretrain-T1** on this, we would not achieve as high of a `T2` AUC as if we had finetuned **Pretrain-T2**. However, it is unlikely we have access to **Pretrain-T2** because existing pretrained models are often biased towards having been trained on, for example, geographically homogenous images [10, 50, 59]. Thus, in our setting in which we are actually able to compare to **Pretrain-T2**, we ask what proportion of the finetuning dataset needs to be `T2` such that the **Pretrain-T1** we actually have can achieve comparable performance. In other words, the amount of change to the finetuning dataset that correlates with having used an entirely different pretrained model to begin with.

CelebA. On our 12 attribute pairs we find that for 128 finetuning samples, manipulating the proportion of `T2` from just 10% to 30% brings the performance of **Pretrain-T1** to that of **Pretrain-T2** in 6 of our 12 cases, and that manipulation on for 1024 finetuning samples does so in 8 of our 12 cases.

Dollar Street and COCO. Applying this insight from CelebA that relatively minor manipulations to the proportion of the dataset from underrepresented subcategories can significantly impact the performance of those subcategories, we now turn to the more complex and realistic object datasets of Dollar Street and COCO.

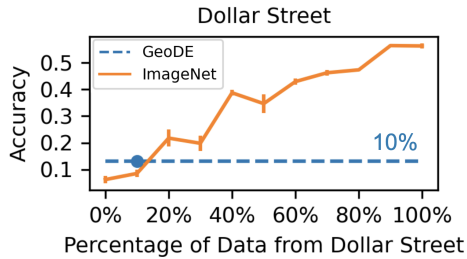We consider the task of recognizing 15 objects in Dollar

Figure 6. The performance of an **ImageNet**-pretrained model [42] at recognizing 15 objects in the geographically diverse Dollar Street [41] dataset, as the proportion of finetuning images from Dollar Street itself (more geographically diverse) and COCO [31] (less so) is manipulated. Only a minor change in finetuning proportion is required for the **ImageNet**-pretrained model to match the performance of a **GeoDE**-pretrained model, which was trained on a dataset curated to be more geographically diverse [40].

Street that have corresponding objects in COCO. Although the object classes are the same between datasets, their visual distribution is different as COCO images largely come from only higher-income regions [10] whereas Dollar Street was collected to be more geographically diverse. Nevertheless COCO images are more plentiful; to simulate this, we consider a finetuning dataset of 128 images where 90% are from COCO and 10% from Dollar Street.

We use two pretrained models, named after the dataset each is trained on: **ImageNet** [42], where the training data is more similar to the COCO distribution, and **GeoDE** [40], which is trained on a newer and more geographically diverse dataset. As expected, finetuning **GeoDE** achieves a higher accuracy (where a prediction is correct if the object is one of the top-5 predicted labels [10]) of 13.1% on Dollar Street compared to finetuning **ImageNet** with 8.4%.

However, what we ultimately want to investigate is how much investing in a better finetuning dataset can help overcome the problem (Fig. 6). Thus, we manipulate the finetuning dataset (simulating the collection of more Dollar Street-like images, while keeping the overall finetuning number the same), and observe that with just 20% rather than 10% of images coming from Dollar Street, **ImageNet** is able to outperform the performance of the **GeoDE** baseline with an accuracy of 21.7%.

6. Discussion

Across both operationalizations of bias, as *spurious correlations* and *underrepresentation*, we find that bias from a pretrained model can propagate into a finetuned one. When bias is in the form of a *spurious correlation*, this is especially likely when the downstream task has a high correlation level with the sensitive attribute, is of low salience, and there is a small number of finetuning samples. For those creating pretrained models, responsibility needs to be taken to clearly document the data the model was trained on, as

well as known biases it may have [13, 33].

However, across both of these conceptualizations of bias, we also find that the distribution of the finetuning dataset has the ability to counteract much of the bias a pretrained model may bring. When the bias we are concerned with is a *spurious correlation*, this entails manipulating the correlation level in the finetuning dataset, and when the bias is *underrepresentation*, this entails manipulating the proportion of positive labels which are of the underrepresented group. Necessarily, all of these manipulations will require additional data collection efforts. However, the efforts are significantly lower than would be required to intervene on the massive pretraining dataset. It is notable that these manipulations allow us to retain most of the performance gains that a biased pretrained base might bring. Acting on these findings will require careful care and consideration in the collection of finetuning datasets, at times even over-correcting for a particular bias that may be present in the downstream test set. Especially given that most settings we investigate are those with smaller finetuning datasets, these being the scenarios where pretrained models are most necessary, we should expect and be willing to put in thoughtful work in curating finetuning datasets [46, 37, 45]. Data collection can often be extractive and violate privacy [25, 38], but there are ways in which it can be done more consensually [40]. Given that the relevant harms are usually context-specific and hard to conceptualize upstream of the downstream task, this lends further support to the notion that the finetuning dataset makes for both a practical and efficient point of intervention in bias mitigation.

6.1. Limitations

There are a number of limitations that qualify the generalizability of our findings. All experiments are conducted on a ResNet50, which uses a convolutional neural network architecture. Prior work validates that a subset of their experiments which are similar to ours have comparable results across model architectures, and we build on this intuition [44]. However, as the size of pretrained models increases dramatically, we leave for future work further exploration. Additionally, we finetune all of the model's weights rather than freezing some.

Most significantly, we only conceive of two possible operationalizations of bias in this work: spurious correlations and underrepresentation. Other types of biases (e.g., stereotypical representations) will likely lead to different transference properties. In our technically defined notions of bias, we have also left out of scope considerations such as NSFW content, privacy issues, and other harms of pretraining datasets [4, 38]. Irrespective of their downstream effects, pretraining on these images are inherently harmful because of impacts such as on the individuals potentially tasked with annotating the data. Ultimately in this work,

we explored very directed interventions on the finetuning dataset to target specific forms of bias. However, likely any downstream task will have many different kinds of bias that are relevant, and potentially even at odds [58]. We leave for future work how to balance potentially conflicting tensions in order to curate more suitable finetuning datasets.

7. Conclusion

Finetuning on top of pretrained models is a powerful way to train models on domains where we have less data. In this work, we conceive of bias as *spurious correlations* or *underrepresentation* and show that biases of either form can transfer from pretrained models to finetuned ones. However, we also affirmatively show that targeted manipulations of the finetuning dataset can counteract this bias transferrance, allowing us to retain performance gains that certain pretrained models with bias may bring, without compromising on fairness concerns. These dual findings indicate that while interventions on the pretrained model to ensure less bias in the features are certainly useful, more effective interventions can be performed by manipulating the finetuning dataset itself. The benefit of manipulations at this juncture is there will also likely be a better understanding of the application-specific harms for which dataset intervention can be targeted. This will require a careful, participatory, and deliberative curation of the finetuning dataset.

Acknowledgements

References

[1] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[2] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. *International Conference on Machine Learning (ICML)*, 2017.

[3] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. *Proceedings of SIGCIS*, 2017.

[4] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv:2110.01963*, 2021.

[5] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 2018.

[7] Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *Association for Computational Linguistics (ACL)*, 2022.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning (ICML)*, 2020.

[9] Kate Crawford and Trevor Paglen. Excavating ai:the politics of training sets for machine learning. *The AI Now Institute, NYU*, 2019.

[10] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? *Conference on Computer Vision and Pattern Recognition workshops (CVPRW)*, 2019.

[11] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. *International Conference on Machine Learning (ICML)*, 2014.

[12] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv:2112.10740*, 2021.

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Hanna Wallach Jennifer Wortman Vaughan, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. *Association for Computational Linguistics (ACL)*, 2021.

[16] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv:2202.08360*, 2022.

[17] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[18] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv:1708.06733*, 2017.

[19] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. *Conference on Human Factors in Computing Systems (CHI)*, 2018.

[20] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *International Conference on Computer Vision (ICCV)*, 2019.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *International Conference on Machine Learning (ICML)*, 2019.

[25] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[26] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *International Conference on Learning Representations (ICLR)*, 2023.

[27] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 2020.

[28] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[29] Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning. *International Conference on Learning Representations (ICLR)*, 2020.

[30] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Ross Girshick Lubomir Bourdev, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)*, 2014.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *International Conference on Computer Vision (ICCV)*, 2015.

[33] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Inioluwa Deborah Raji Elena Spitzer, and Timnit Gebru. Model cards for model reporting. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

[34] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[35] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. pages 8024–8035, 2019.

[37] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, and Alex Hanna Emily Denton. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2021.

[38] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[39] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[40] Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. *arXiv:2301.02560*, 2023.

[41] William A. Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *NeurIPS Datasets and Benchmarks*, 2022.

[42] 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.

[43] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *International Conference on Machine Learning (ICML)*, 2020.

[44] Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. When does bias transfer in transfer learning? *arXiv:2207.02842*, 2022.

[45] Nithya Sambasivan, Shivani Kapania, H. Highfill, Diana Akrong, Praveen K. Paritosh, and Lora Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. *Conference on Human Factors in Computing Systems (CHI)*, 2021.

[46] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *ACM Conference on Human-Computer Interaction (CSCW)*, 2021.

[47] Morgan Klaus Scheuerman, Aaron Jiang, Katta Spiel, and Jed R. Brubaker. Revisiting gendered web forms: An evaluation of gender inputs with (non-)binary people. *Conference on Human Factors in Computing Systems (CHI)*, 2021.

[48] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. A step toward more inclusive people annotations for fairness. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

[49] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[50] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *NIPS 2017 Workshop on Machine Learning for the Developing World*, 2017.

[51] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[52] Kirill Sirotkin, Pablo Carballeira, and Marcos Escudero-Viñolo. A study on the distribution of social biases in self-supervised learning visual models. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[53] Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted

datasets. *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[54] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *Workshop on Gender Bias in Natural Language Processing at NAACL*, 2022.

[55] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[56] Ryan Steed, Swetasudha Panda, and Michael Wick Ari Kobren. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. *Association for Computational Linguistics (ACL)*, 2022.

[57] Guillermo Valle-Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *International Conference on Learning Representations (ICLR)*, 2019.

[58] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[59] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision (IJCV)*, 2022.

[60] Angelina Wang and Olga Russakovsky. Directional bias amplification. *International Conference on Machine Learning (ICML)*, 2021.

[61] 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. *International Conference on Computer Vision (ICCV)*, 2019.

[62] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *European Conference on Computer Vision (ECCV)*, 2014.

[63] Dora Zhao, Jerone T.A. Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. *International Conference on Machine Learning (ICML)*, 2023.

[64] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. *International Conference on Computer Vision (ICCV)*, 2021.

[65] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.