# Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations

Tianlu Wang[1], Jieyu Zhao[2], Mark Yatskar[3], Kai-Wei Chang[2], Vicente Ordonez[1]
[1]University of Virginia, [2]University of California Los Angeles,
[3]Allen Institute for Artificial Intelligence
tianlu@virginia.edu, jyzhao@cs.ucla.edu, marky@allenai.org,
kwchang@cs.ucla.edu, vicente@virginia.edu

## Abstract

*In this work, we present a framework to measure and mitigate intrinsic biases with respect to protected variables –such as gender– in visual recognition tasks. We show that trained models significantly amplify the association of target labels with gender beyond what one would expect from biased datasets. Surprisingly, we show that even when datasets are balanced such that each label co-occurs equally with each gender, learned models amplify the association between labels and gender, as much as if data had not been balanced! To mitigate this, we adopt an adversarial approach to remove unwanted features corresponding to protected variables from intermediate representations in a deep neural network – and provide a detailed analysis of its effectiveness. Experiments on two datasets: the COCO dataset (objects), and the imSitu dataset (actions), show reductions in gender bias amplification while maintaining most of the accuracy of the original models.*

## 1. Introduction

While visual recognition systems have made great progress toward practical applications, they are also sensitive to spurious correlations and often depend on these erroneous associations. When such systems are used on images containing people, they risk amplifying societal stereotypes by over associating protected attributes such as gender, race or age with target predictions, such as object or action labels. Known negative outcomes have included representation harms (e.g., male software engineers are being over-represented in image search results [11]), harms of opportunity, (e.g., facial recognition is not as effective for people with different skin tones [3]), to life-threatening situations (e.g., recognition rates of pedestrians in autonomous vehicles are not equally accurate for all groups of people [32]).

In this paper we study gender bias amplification: the effect that trained models exaggerate gender stereotypes that are present in the training data. We focus on the tasks of recognizing objects in the COCO dataset [16] and actions in the imSitu dataset [36], where training resources exhibit gender skew and models trained on these datasets exhibit bias amplification [39].[1] In an effort to more broadly characterize bias amplification, we generalize existing measures of bias amplification. Instead of measuring the similarity between training data and model prediction distributions, we compare the predictability of gender from ground truth labels (*dataset leakage*, Figure 1 on the top) and model predictions (*model leakage*, Figure 1 on the bottom). Each of these measures is computed using a classifier that is trained

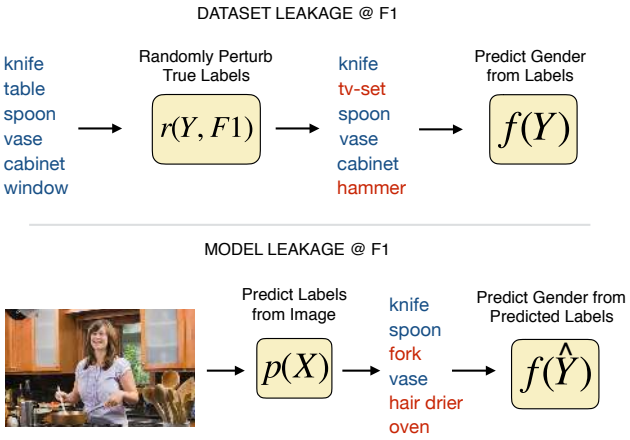


Figure 1. On the top we illustrate our newly introduced concept of *Dataset Leakage* which measures the extent to which gender –or more generally a protected variable– can be inferred from randomly perturbed ground-truth labels. On the bottom we illustrate our concept of *Model Leakage* which measures the extent to which gender can be inferred from the outputs of a model. A model amplifies bias if model leakage exceeds dataset leakage.

[1]For example women are represented as cooking twice as often as men in imSitu, but after models are trained and evaluated on similarly distributed data, they predict cooking for women three times as often as men.

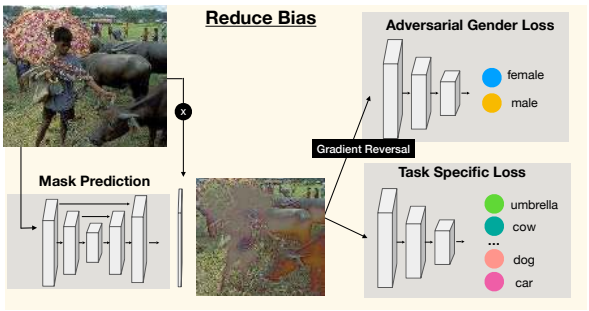


Figure 2. In our bias mitigation approach, we learn a task-specific model with an adversarial loss that removes features corresponding to a protected variable from an intermediate representation in the model – here we illustrate our pipeline to visualize the removal of features in image space through an auto-encoder network.

to predict gender from either ground truth labels or models predictions. We say a model exhibits bias amplification if it leaks more information about gender than a classifier of equivalent accuracy whose errors are only due to chance.

Our new leakage measures significantly expand the types of questions we can ask about bias amplification. While previously it was shown that models amplify bias when they are required to predict gender alongside target variables [39], our empirical findings indicate that when models are not trained to predict gender, they also amplify gender bias. Surprisingly, we find that if we additionally balance training data such that each gender co-occurs equally with each target variable, models amplify gender bias as much as in unbalanced data! This strongly argues that naive attempts to control for protected attributes when collecting datasets will be ineffective in preventing bias amplification.

We posit that models amplify biases in the data balanced setting because there are many gender-correlated but unlabeled features that cannot be balanced directly. For example in a dataset with equal number of images showing men and women cooking, if *children* are unlabeled but co-occur with the *cooking* action, a model could associate the presence of children with *cooking*. Since children co-occur with women more often than men across all images, a model could label women as *cooking* more often than we expect from a balanced distribution, thus amplifying gender bias.

To mitigate such unlabeled spurious correlations, we adopt an adversarial debiasing approach [34, 2, 38, 6]. Our goal is to preserve as much task specific information as possible while eliminating gender cues either directly in the image or intermediate convolutional representations used for classification. As seen in Figure 2, models are trained adversarially to trade off a task-specific loss while trying to create a representation from which it is not possible to predict gender. For example, in Figure 3 in the bottom right image, our method is able to hide regions that indicate the gender of the main entity while leaving enough information to determine that *she* is weight lifting.

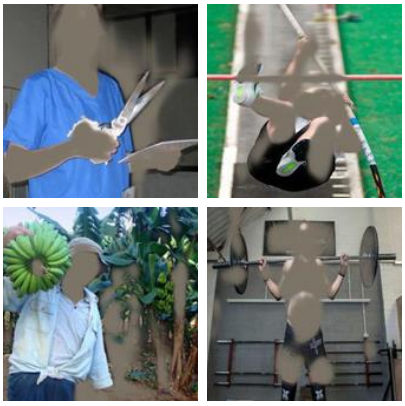

Figure 3. Images after adversarial removal of gender when applied to the image space. The objective was to preserve information about objects and verbs, e.g. scissors, banana (COCO) or vaulting, lifting (imSitu) while removing gender correlated features.

Evaluation of our adversarial debiased models show that they are able to make significantly better trade-offs between task accuracy and bias amplification than other methods. We consider strong baselines that include masking or blurring out entities by having access to ground truth mask annotations for people in the images. We also propose a baseline that simply adds noise to intermediate representations – thus reducing the ability to predict gender from features, but often at a significant compromise in task accuracy. Of all methods considered, only adversarial debiasing provided a better trade-off compared to randomizing model predictions, and we were able to reduce bias amplification by 53-67% while only sacrificing 1.2 - 2.2 points in accuracy.

## 2. Related Work

Recently, researchers have demonstrated that machine learning models tend to replicate societal biases present in training datasets. Concerns have been raised for applications such as recommender systems [35], credit score prediction [9], online news [24], and others [11] and in response various approaches have been proposed to mitigate bias [1, 10]. However, most previous work deals with issues of resource allocation [5, 7] where the focus is on calibrating predictions. Furthermore, works in this domain often assume protected variables are explicitly specified as features, making the goal of calibration more clearly defined. However in visual recognition, representations for protected attributes are automatically inferred from raw data.

There are also works addressing biases in images [25, 39, 27, 3, 20, 4]. Zhao et al [39] reduces bias in structured prediction models where gender is one of the target variables. Burns et al [4] attempts to calibrate gender predictions of a captioning system by modifying the input image. In contrast, our work focuses on models that are not aimed at predicting gender, which is a more common scenario. Calibration methods would not be effective to debias

in our proposed setup, as gender is not one of the outputs.

Our work is motivated by previous efforts on adversarial debiasing in various other tasks and domains [38, 2, 34, 6, 40, 8]. We provide further details about this family of methods in the body of the paper, and adopt this framework for debiasing the intermediate results of deep neural networks. Our work advances the understanding of this area by exploring what parts of deep representations are the most effective to debias under this approach, and we are the first to propose a way to visualize such debiased representations.

Issues of dataset bias have been addressed in the past the computer vision community [30, 12, 29]. Torralba and Efros [30] showed that it was possible to identify the source dataset given image samples for a wide range of standard datasets, and [12] addresses this issue by learning shared parameters across datasets. More recently, Tommasi et al [29] provided a fresher perspective on this issue using deep learning models. There are strong connections with these prior works when dataset source is to be taken as a protected variable. Our notion of bias is more closely related to the notion of bias used in the fairness in machine learning literature, where there is protected variable (e.g. gender) for which we want to learn unbiased representations (e.g. [37]).

In terms of evaluation, researchers have proposed different measurements for quantifying fairness [9, 15, 5]. In contrast, we try to reduce bias in the feature space. We adopt and further develop the idea of *leakage* as an evaluation criteria, as proposed by Elazar and Goldberg [6] to debias text representations. We significantly expand the *leakage* formulation and propose *dataset leakage*, and *model leakage* as measures of bias in learned representations.

Building models under fairness objectives is also related to feature disentangling methods [28, 22, 17, 18, 19]. However, most research in this domain has focused on facial analysis – where there is generally more well aligned features. This general area of work is also related to efforts in building privacy preserving methods [31, 26, 33, 13], where the objective is to obfuscate the input while still being able to perform a recognition task. In contrast, in fairness methods, there is no requirement to obfuscate the inputs, and in particular the method proposed in this paper is most effective when applied to intermediate feature representations.

## 3. Leakage and Amplification

Many problems in computer vision inadvertently reveal demographic information in images. For example, in COCO, images of plates contain significantly more women than men. If a model predicts that a plate is in the image, we can infer there is likely a woman as well. We refer to this notion as *leakage*. In this section, we present formal definitions of leakage for a dataset and models, and a measure for quantifying bias amplification as summarized in Figure 1.

**Dataset Leakage:** Given an annotated dataset $\mathcal{D}$ containing instances $(X_i, Y_i, g_i)$, where $X_i$ is an image annotated with a set of task-specific labels $Y_i$ (e.g., objects), and a protected attribute $g_i$ (e.g., the image contains a male/female person)[2], we say that a particular annotation $Y_i$ leaks information about $g_i$ if there exists a function $f$ such that $g_i \approx f(Y_i)$. We refer to this $f$ as an *attacker* as it tries to reverse engineer information about protected attributes in the input image $X_i$ only from its task-specific labels $Y_i$. To measure leakage across a dataset, we train such an attacker and evaluate it on held out data. The performance of the attacker, the fraction of instances in $\mathcal{D}$ that leak information about $g_i$ through $Y_i$, yields an estimate of leakage:

$$\lambda_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum\nolimits_{(Y_i, g_i) \in \mathcal{D}} \mathbb{1}[f(Y_i) == g_i],$$

where $\mathbb{1}[\cdot]$ is the indicator function. We extend this definition of leakage to assess how much gender is revealed at different levels of accuracy, where errors are due entirely to chance. We define dataset leakage at a performance $a$ by perturbing ground truth labels, with some function $r(Y_i, a)$, such that the overall accuracy of the changed labels with respect to the ground truth achieves an accuracy $a$:

$$\lambda_D(a) = \frac{1}{|\mathcal{D}|} \sum\nolimits_{(Y_i, g_i) \in \mathcal{D}} \mathbb{1}[f(r(Y_i, a)) == g_i],$$

This allows us to measure the leakage of a model whose performance is $a$ and whose mistakes cannot be attributed to systematic bias. Across all experiments, we use F1 as the performance measure, and $\lambda_D = \lambda_D(1.0)$, by definition.

**Model Leakage:** Similar to dataset leakage, we would like to measure the degree a model, $M$ produces predictions, $\hat{Y}_i = M(X_i)$, that leak information about the protected variable $g_i$. We define model leakage as the percentage of examples in $\mathcal{D}$ that leak information about $g_i$ through $\hat{Y}_i$. To measure prediction leakage, we train a different attacker on $\hat{Y}_i$ to extract information about $g_i$:

$$\lambda_M(a) = \frac{1}{|\mathcal{D}|} \sum\nolimits_{(\hat{Y}_i, g_i) \in \mathcal{D}} \mathbb{1}[f(\hat{Y}_i) == g_i)],$$

where $f$ is a attacker function trained to predict gender from the outputs of model $M$ which has an accuracy score $a$.

**Bias Amplification:** Formally, we define the bias amplification of a model $p$, as the difference between the *model leakage* and the *dataset leakage* at the same accuracy $a$.

$$\Delta = \lambda_M(a) - \lambda_D(a) \tag{1}$$

Note that $\lambda_D(a)$ measures the leakage of an ideal model which achieves a performance level $a$ but only makes mistakes randomly, not due to systematic bias. A model with $\Delta$

[2]In this paper, we assume gender as binary due to the available annotations, but the work could be extended to non-binary, as well as a broader set of protected attributes, such as race or age.

larger than zero leaks more information about gender than we would expect even from simply accomplishing the task defined by the dataset. This represents a type of amplification on the reliance on protected attributes to accomplish the prediction task. In Eq. (1), $a$ could be any performance measurement but we use F1 score throughout our experiments. We show later in Section 4 that all models we evaluated leak more information than we would expect and even leak information when the dataset does not.

**Creating an Attacker:** Ideally, the attacker should be a Bayes optimal classifier, which makes the best possible prediction of $g$ using $Y$. However, in practice, we need to train a model to do prediction for every model, and we use a deep neural network to do so. Yet, we are not guaranteed to obtain the best possible function for mapping $y$ to $g$. Thus, it is important to consider the reported leakage as a lower bound on true leakage. In practice, we find that we can robustly estimate $f$ (see Section 4: Attacker Learning is Robust).

# 4. Bias Analysis

In this section we summarize our findings that both imSitu and COCO leak information about gender. We show that models trained on these datasets leak more information than would be expected (1) when models are required to predict gender through a structured predictor that jointly predicts labels and gender, (2) when models are required to predict only labels, and (3) even when not predicting gender and datasets were balanced such that each gender co-occurs equally with target labels. Table 1 summarizes our results.

## 4.1. Experiment Setup

We consider two tasks: (1) multi-label classification in the COCO dataset [16], including the prediction of gender, and (2) imSitu activity recognition, a multi-classification task for people related activities.

**Datasets:** We follow the setup of existing work for studying bias in COCO and imSitu [39], deriving gender labels from captions and "agent" annotations respectively. For the purpose of analysis, we exclude "person" category and only use images containing people. We have 22826, 5367, 5473 and 24301, 7730, 7669 images in the training, validation and testing set for COCO and imSitu respectively.

**Models:** For both object and activity recognition, we use a standard ResNet-50 pretrained on Imagenet (ILSVRC) as the underlying model by replacing the last linear layer. We also consider the Conditional Random Field (CRF) based model in [39] when predicting gender jointly with target variables. Attackers are a 4-layer multi-layer perceptron (MLP) with BatchNorm and LeakyReLU in between.

**Metrics:** We use mAP, or the mean across categories of the area under the precision-recall curve, and F1 score for both tasks by using the discrete output predictions of the model.

**Computing Leakage:** Model leakage is predicted from pre-activation logits while dataset leakage is predicted from binary labels. Attackers are trained and evaluated with an equal amount images of men and women.

**Training Details:** All models are developed and evaluated on the same dev and test sets from the original split. We optimize using Adam [14] with a learning rate of $10^{-4}$ and a minibatch size of 32 to train the linear layers for classification. We then fine-tune the model with a learning rate of $5 \times 10^{-6}$. We train all attackers for 100 epochs with a learning rate of $5 \times 10^{-5}$ and a batch size of 128, keeping the snapshot that performs best on the dev set.

## 4.2. Results

**Dataset Leakage:** Dataset leakage measures the degree to which ground truth labels can be used to estimate gender. The rows corresponding to "original CRF" in Table 1 summarize dataset leakage in imSitu and COCO ($\lambda_D$). Both datasets leak information: the gender of a main entity in the image is extractable from ground truth annotations 67.72% and 68.26% for COCO and imSitu, respectively.

**Bias Amplification:** Bias amplification ($\Delta$) captures how much more information is leaked than what we expect from a similar model which makes mistakes entirely due to chance. Dataset leakage needs to be calibrated with respect to model performance for computing bias amplification. To do so, we randomly flip ground truth labels to reach various levels of accuracy. Figure 4 shows dataset leakage at different performance levels in COCO and imSitu. The relationship between F1 and leakage is roughly linear. In Table 1, we report adjusted leakage for models at appropriate levels ($\lambda_D(\text{F1})$). Finally, bias amplification ($\Delta$) can be computed by taking the difference between adjusted dataset leakage ($\lambda_D(\text{F1})$) and model leakage ($\lambda_M(\text{F1})$).

Models trained on standard splits of both COCO and imSitu that jointly predict gender and target labels (the original rows in Table 1), all leak significantly more gender information than we would expect by chance. Surprisingly, imSitu is more gender balanced than COCO but actually leaks significantly more information than models trained on COCO. When models are no longer required to predict gender, they leak less information than before but still more than we would expect (the *no gender* rows in Table 1).

**Alternative Data Splits:** It is possible to construct datasets which leak less through subsampling. We obtain splits more balanced in male and female co-occurrences with labels by imposing the constraint that neither gender occurs more frequently with any output label by a ratio greater than $\alpha$:

$$\forall y : 1/\alpha < \#(m, y)/\#(w, y) < \alpha, \tag{2}$$

where $\#(m, y)$ and $\#(w, y)$ are the number of occurrences of *men* with label $y$ and of *women* with label $y$ respectively. Enforcing this constraint in imSitu is trivial because each

| dataset | Statistics split | #men | #women | Leakage $\lambda_D$ | $\lambda_M$(F1) | $\lambda_D$(F1) | Performance $\Delta$ | mAP | F1 |
|---|---|---|---|---|---|---|---|---|---|
| COCO [16] | original CRF | 16,225 | 6,601 | $67.72 \pm 0.31$ | $73.20 \pm 0.59$ | 60.35 | 12.85 | 57.77 | 52.52 |
| | no gender | 16,225 | 6,601 | $67.72 \pm 0.31$ | $70.46 \pm 0.36$ | 60.53 | 9.93 | 58.23 | 53.75 |
| | ($\alpha = 3$) | 10,876 | 6,598 | $62.00 \pm 0.98$ | $67.78 \pm 0.29$ | 57.50 | 10.28 | 57.04 | 52.60 |
| | ($\alpha = 2$) | 8,885 | 6,588 | $56.77 \pm 1.45$ | $64.45 \pm 0.56$ | 54.72 | 9.73 | 56.21 | 51.95 |
| | ($\alpha = 1$) | 3,078 | 3,078 | $53.15 \pm 1.10$ | $63.22 \pm 1.11$ | 52.85 | 10.37 | 48.23 | 42.89 |
| imSitu [36] | original CRF | 14,199 | 10,102 | $68.26 \pm 0.31$ | $78.43 \pm 0.26$ | 56.58 | 21.85 | 41.83 | 40.75 |
| | no gender | 14,199 | 10,102 | $68.26 \pm 0.31$ | $76.93 \pm 0.20$ | 56.46 | 20.47 | 41.02 | 40.11 |
| | ($\alpha = 3$) | 11,613 | 9,530 | $68.11 \pm 0.55$ | $75.79 \pm 0.49$ | 55.98 | 19.81 | 39.20 | 37.64 |
| | ($\alpha = 2$) | 10,265 | 8,884 | $68.15 \pm 0.32$ | $75.46 \pm 0.32$ | 55.74 | 19.72 | 37.53 | 36.41 |
| | ($\alpha = 1$) | 7,342 | 7,342 | $53.99 \pm 0.69$ | $74.83 \pm 0.34$ | 53.20 | 21.63 | 34.63 | 33.94 |

Table 1. In this table we show for different splits in COCO and imSitu, (1) $\lambda_D$, dataset leakage or the accuracy obtained by predicting gender from ground truth annotations, showing that our data balancing approach successfully achieves significantly reducing this type of leakage (2) $\lambda_M$(F1), model leakage or the accuracy obtained by a model trained to predict gender on the outputs of a model trained on the target task, the last two columns show the mAP and F1 score of the model, and (3) $\lambda_D$(F1), dataset leakage at a certain performance leverl, or the leakage of a model with access to ground truth annotations but with added noise so that its accuracy matches that of a model trained on this data, i.e. same F1 as shown in the last column. (4) $\Delta$, bias amplification, the difference between model leakage and dataset leakage at the same performance level, indicating how much more leakage the model is exhibiting over chance.

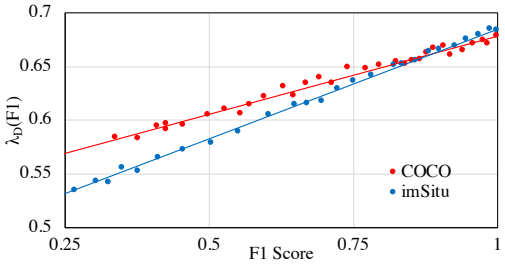


Figure 4. Dataset leakage in COCO and imSitu as function of F1 score. Ground truth labels were randomly flipped to simulate a method that performs at different levels of F1 score. We refer to this accuracy adjusted leakage as $\lambda_D(F1)$), or the amount we would expect a method to leak given its performance level.

image is only annotated with one verb: we simply sample the over-represented gender to pass the above constraints. For COCO, we heuristically enforce this constraint since each image contains multiple object annotations. We try to make every object satisfy this constraint one at a time, removing images having less objects. We iterate through all objects until this process converges and all objects satisfy the constraint. We create splits for $\alpha \in \{3, 2, 1\}$.[3]

Table 1 rows $\alpha = \{3, 2, 1\}$ summarize results for re-balancing data with respect to gender. As we expect, decreasing values of $\alpha$ yields smaller datasets with less dataset leakage but worse predictors because there is less data. Yet model leakage does not reduce as quickly as dataset leakage, resulting in nearly no change in bias amplification. In fact, when there is nearly no dataset leakage, models still leak information. Likely this is because it is impossible to balance *unlabeled* co-occurring features with gender (e.g. COCO only has annotations for 80 objects) and the models still rely on these features to make predictions. In summary, **balancing the co-occurance of gender and target labels does not reduce bias amplification in a meaningful way.**

[3]Practically satisfying $\alpha = 1$ is in-feasible, but our heuristic is able to find a set where $\alpha = 1.08$.

| Attacker | $\lambda_M$ |
|---|---|
| 1 layer , ———- , all data | $68.82 \pm 0.35$ |
| 2 layer , 100 dim , all data | $70.83 \pm 0.58$ |
| 2 layer , 300 dim , all data | $71.03 \pm 0.52$ |
| 4 layer , 300 dim , all data | $70.46 \pm 0.36$ |
| 4 layer , 300 dim , 75% data | $69.93 \pm 0.51$ |
| 4 layer , 300 dim , 50% data | $69.89 \pm 0.98$ |
| 4 layer , 300 dim , 25% data | $68.54 \pm 1.10$ |

Table 2. Varying attacker architecture and training data when estimating model leakage on the original COCO. The leakage estimate is robust to significant changes, showing that estimation of leakage with our adversaries is largely easy and stable.

**Attacker Learning is Robust:** Measuring leakage relies on being able to consistently estimate an attacker. To verify that leakage estimates are robust to different architectures and data settings on the attacker side, we conduct an ablation study in Table 2. We vary the attacker architecture and the amount of training data to measure model leakage ($\lambda_M$). Except an attacker with 1-layer, none of the others vary in their estimation of leakage by more than 2 points.

## 5. Adversarial Debiasing

In this section we show the effectiveness of a method for reducing leakage through training with an auxiliary adversarial loss which effectively removes gender information from intermediate representations. We additionally propose a way to visualize the effects of this approach on the input space, to inspect the type of information being removed.

### 5.1. Method Overview

We propose a simple formulation for reducing the amount of leakage in a model, summarized in Figure 2. We hypothesize that models leak extra information about protected attributes because the underlying representation is overly sensitive to features related to those attributes. As such, we encourage models to build representations from which protected attributes can not be predicted.

Our methods rely on the construction of a *critic*, $c$, which attempts to predict protected information from an intermediate representation, $h_i$ for a given image $X_i$, of a predictor, $p$. The critic attempts to minimize a loss over the amount of information it can extract:

$$\sum\nolimits_{(h_i,g_i)\in\mathcal{D}} L_c(c(h_i), g_i),$$

while the predictor tries to minimize its loss over the task specific predictions while increasing the critic's loss:

$$L_p = \sum\nolimits_{(X_i,h_i,Y_i)\in\mathcal{D}} [L(p(X_i), Y_i) - \lambda L_c(c(h_i), g_i)].$$

In both cases, $L$ is the cross-entropy loss, and when optimizing $L_p$ we do not update $c$, and trade-off task performance with sensitivity to protected attributes with $\lambda$.

We also experiment with optimizing the adversarial loss on the input feature space by leveraging an encoder-decoder model that auto-encodes the input image $X_i$. In order to accomplish this goal, we add an additional loss with a weight parameter $\beta$ to the predictor as follows:

$$L_p = \sum\nolimits_i \left[\beta|X_i - \hat{X}_i|_{\ell_1} + L(p(\hat{X}_i), Y_i) - \lambda L_c(c(\hat{X}_i), g_i)\right]$$

Where $\hat{X}_i = M_i \cdot X_i$, which is the original image element-wise multiplied with a mask $M_i$ generated by an encoder-decoder bottleneck network with input $X_i$. So the first term is encouraging the mask to maintain the information in the original image, the second term is trying to obtain correct task-specific predictions from the masked input, and the third term is adversarially trying to obscure gender by modifying the mask. This is similar to the proposed experiment in Palacio et al [21] where instead, the outputs of an autoencoder are directly fed to a convolutional neural network trained to recognize objects in order to interpret the patterns learned by the network. In contrast, our objective is to visualize what the adversary learned to obfuscate while trying to preserve accurate results.

### 5.2. Implementation Details

We first train the classification layers (linear classifiers) with $10^{-4}$ as learning rate and a batch size of 32 until the performance plateaus. We then incorporate the adversarial loss, and fine-tune the model end-to-end using a learning rate $5{\times}10^{-6}$. Before activating the adversarial loss, we first train the gender classification branch so that its gradients provide useful guidance for feature removal during adversarial training. In every batch, we sample the same amount of male and female images for training this adversary.

### 5.3. Models

**Adversarial Methods** We consider three different types of adversaries which try to remove leakage at different stages in a ResNet-50 classification network.

- **adv @ image**, or removing gender information directly at the image. We use U-Net [23] as our encoder-decoder network to predict a mask $M_i$. The original image is point-wise multiplied with this mask and then fed to two branches. The first branch is a ResNet-18 which attempts to detect gender (the adversary) and the second branch is a ResNet-50 for classifying the target categories.
- **adv @ conv4**, removes gender information from an intermediate hidden representation of ResNet-50 (on the 4th convolutional block). We use an adversary with 3 convolutional layers and 4 linear layers.
- **adv @ conv5**, removes gender information from the final convolutional layer of ResNet-50. We use a linear adversary which takes as input a vectorized form of the output feature map and uses a 4-layer MLP for classification.

**Baselines:** We consider several alternatives to adversarial training to reduce leakage, including some that have access to face detectors and ground truth segment annotations.

- **Original**: The basic recognition model, trained on the original data, without any debiasing attempt.
- **Randomization**: Adding Gaussian noise at increasing magnitudes to the pre-classification embedding layer of the original model. We expect larger noise to reduce more leakage while preventing the model from effectively classifying images.
- **Alternative Datasets**: We also consider constructing alternative data splits through downsampling approaches that reduce dataset leakage. We refer to this alternative data splits as $\alpha = 1, 2, 3$, as defined in section 4.2.
- **Blur**: Consists of blurring people in images when ground truth segments are available (COCO only).
- **Blackout - Face**: Consists of blacking out the faces in the images using a face detector.
- **Blackout - Segm**: Consists of blacking out people in images when ground truth segments are available (COCO only). This aggressively removes features such as skin and clothing. It may also obscure objects with which people are closely interacting with.
- **Blackout - Box**: Consists of blacking out people using ground truth bounding boxes (COCO and imSitu). This removes large regions of the image around people, likely removing many objects and body pose cues.

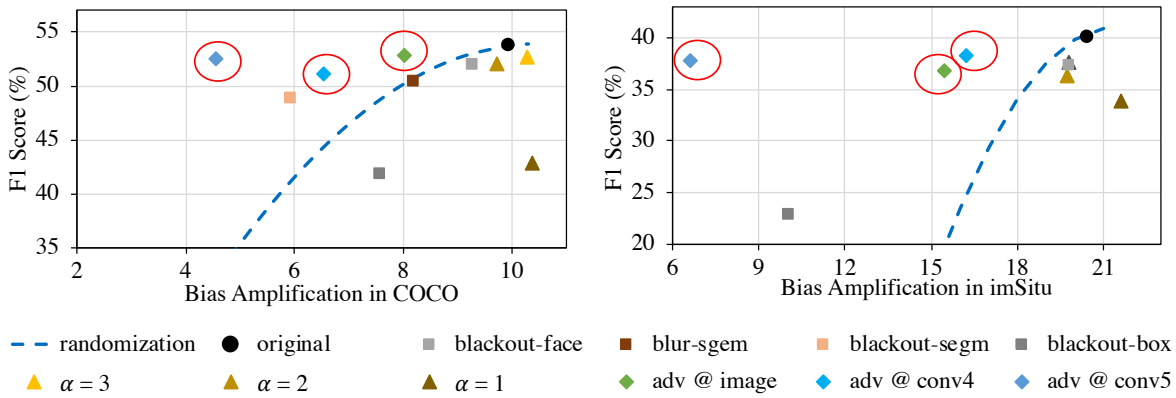


Figure 5. Bias amplification as a function of F1 score on COCO object classification and imSitu action recognition. Models in the top left have low leakage and high F1 score. The blue dashed line indicates bias and performance of adding progressively more noise to the original model representation. Our adversarial methods (circled) are the ones which make a better trade-off between performance and bias amplification than randomization and other baselines.

| | Leakage | | Performance | | |
|---|---|---|---|---|---|
| | $\lambda_M$(F1) | $\lambda_D$(F1) | $\Delta$ | mAP | F1 |
| original CRF | 73.20 | 60.35 | 12.85 | 57.77 | 52.53 |
| CRF + RBA | 73.31 | 60.16 | 13.15 | 56.46 | 51.28 |
| CRF + adv | 65.00 | 60.19 | **4.81** | 56.68 | 51.48 |

Table 3. Model Leakage and performance trade-offs for RBA (Reducing Bias Amplification, proposed in [39]) and our adversarial training methods. We adopt the CRF based model [39] to predict COCO objects as well as the gender. Our method reduces more than 60% bias amplification while RBA fails to do so.

| | Leakage | | Performance | | |
|---|---|---|---|---|---|
| | $\lambda_M$(F1) | $\lambda_D$(F1) | $\Delta$ | mAP | F1 |
| original | 70.46 | 60.53 | 9.93 | **58.23** | **53.75** |
| blackout-face | 69.53 | 60.24 | 9.29 | 55.93 | 51.81 |
| blur-segm | 68.19 | 59.99 | 8.20 | 55.06 | 50.26 |
| blackout-segm | 65.72 | 59.76 | 5.96 | 53.78 | 48.72 |
| blackout-box | 64.00 | 58.71 | 5.29 | 47.42 | 41.81 |
| adv @ image | 68.49 | 60.47 | 8.02 | 56.14 | 52.82 |
| adv @ conv4 | 66.66 | 60.12 | 6.54 | 55.18 | 51.08 |
| adv @ conv5 | 64.92 | 60.35 | **4.57** | 56.35 | 52.54 |
| ($\alpha = 1$) | 63.22 | 52.85 | 10.37 | 48.23 | 42.89 |
| adv @ conv5 | **54.91** | 52.40 | **2.51** | 43.71 | 38.98 |

Table 4. Model leakage and performance trade-offs for different baselines (rows 1-5) and our adversarial training methods (rows 6-8) on COCO object classification. Our methods make significantly better trade-offs than baselines, even improving on methods which use ground truth detection and segmentation. Applying adversarial training on balanced dataset reaches lowest model leakage (54.91) and bias amplification (2.51).

## 5.4. Quantitative Results

Table 4 and Table 5 summarize our results. Adversarially trained methods offer significantly better trade-offs between leakage and performance than any other method. We are able to reduce model leakage by over 53% and 67% on COCO and imSitu respectively, while suffering only 1.21 and 2.26 F1 score degradation. Furthermore, no one class disproportionately suffers after our method (See Figure 7). We also compare our method with RBA [39], a debiasing algorithm proposed to maintain the similarity between the training data and model predictions. As shown in Table 3, the original CRF model predicts gender and objects, RBA fails to have reduce bias amplification. Figure 5 further highlights that our methods are making extremely favorable trade-offs between leakage and performance, even when compared to methods that blur, black-out, or completely remove people from the images using ground truth segment annotations. Adversarial training is the only method that consistently improves upon simply adding noise to the model representation before prediction (the blue curves).

| | Leakage | | Performance | | |
|---|---|---|---|---|---|
| | $\lambda_M$(F1) | $\lambda_D$(F1) | $\Delta$ | mAP | F1 |
| original | 76.93 | 56.46 | 20.47 | **41.02** | **40.11** |
| blackout-face | 75.69 | 55.91 | 19.78 | 38.22 | 37.29 |
| blackout-box | 63.14 | 53.06 | 10.08 | 21.76 | 22.75 |
| adv @ image | 71.32 | 55.83 | 15.49 | 36.90 | 36.88 |
| adv @ conv4 | 72.39 | 56.15 | 16.24 | 38.81 | 38.35 |
| adv @ conv5 | 62.65 | 56.02 | **6.63** | 38.91 | 37.85 |
| ($\alpha = 1$) | 74.83 | 53.20 | 21.63 | 34.63 | 33.94 |
| adv @ conv5 | **57.49** | 52.85 | **4.64** | 30.78 | 30.37 |

Table 5. Model leakage and performance trade-offs for different baselines (rows 1-3) and our adversarial training methods (rows 4-6) on imSitu activity recognition. Our methods make significantly better trade-offs than baselines. Applying adversarial training on balanced dataset reaches lowest model leakage (57.49) and bias amplification (4.64).

## 5.5. Qualitative Results

While adversarial removal works best when applied to representations in intermediate convolutional layers. In or-

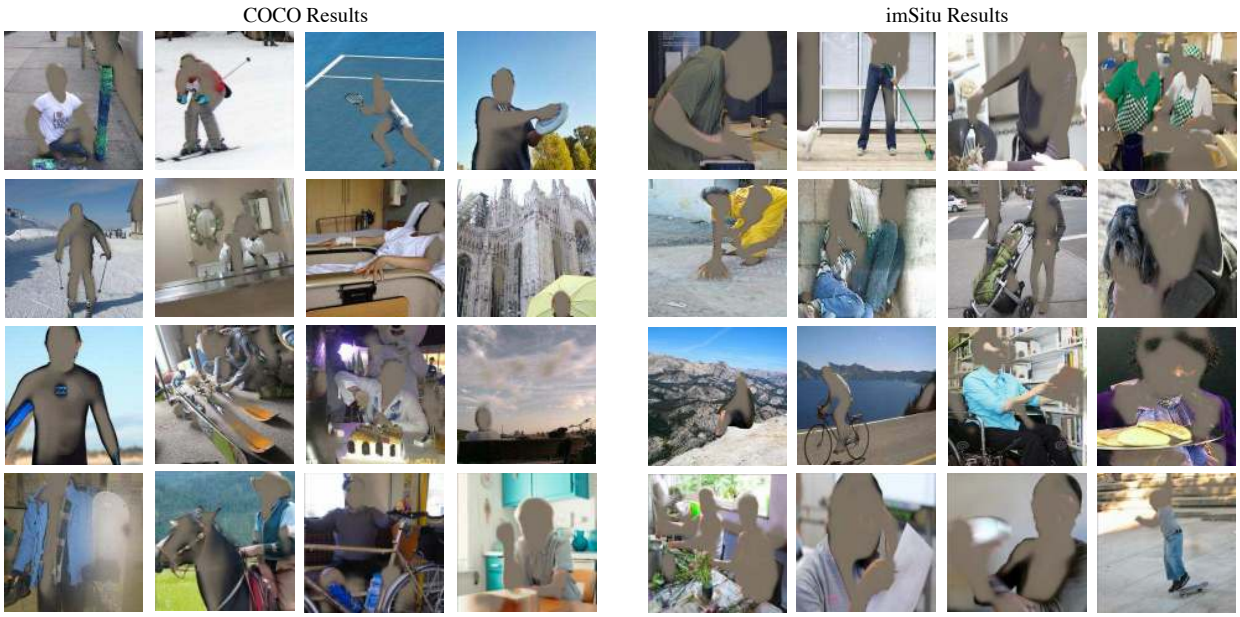


Figure 6. Images after adversarial removal of gender in image space by using a U-Net based autoencoder as inputs to the recognition model. While people are clearly being obscured from the image, the model selectively chooses to obscure only parts that would reveal gender such as faces but tries to keep information that is useful to recognize objects or verbs. 1st row: WWWM MMWW; 2nd row: MWWW WMWW; 3rd row: MMMW MMWM; 4th row: MMMW WWMM. W: woman; M: man.

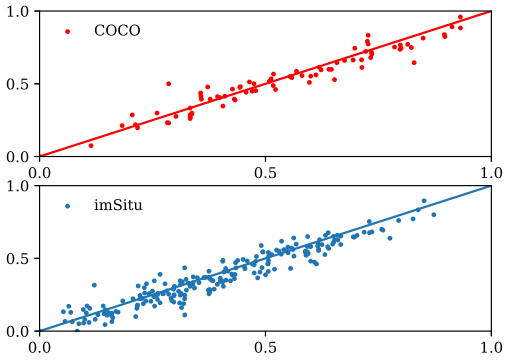


Figure 7. Performance change of every object/verb. X axis: F1 score before debiasing. Y axis: F1 score after debiasing. Across two datasets, most of objects/verbs are very close to the solid line ($y = x$), showing that no one class is disproportionately affected.

der to obtain interpretable results, we apply gender removal in the image space and show results in Fig. 6. In some instances our method removes the entire person, in some instances only the face, in other cases clothing, and garments that might be strongly associated with gender. Our approach learns to selectively obscure pixels enough to make gender prediction hard but leaving sufficient information to predict other things, especially objects that need to be recognized such as *frisbee*, *bench*, *ski*, as well as actions such as *cooking*, *biking*, etc. This is in contrast to our strong baselines that remove the entire person instances using ground-truth segmentation masks. A more sensible compromise is learned through the adversarial removal of gender without the need for segment-level supervision.

## 6. Conclusion

We introduced *dataset leakage*, and *model leakage* as measures of the encoded bias with respect to a protected variable in either datasets or trained models. We demonstrated that models amplify the biases in existing datasets for tasks that are not related to gender recognition. Moreover, we show that balanced datasets do not lead to unbiased predictions and that more fundamental changes in visual recognition models are nedeed. We also demonstrated an adversarial approach for the removal of features associated with a protected variable from the intermediate representations learned by a convolutional neural network. Our approach is superior to applying various forms of random perturbations in the representations, and to applying image manipulations that have access to significant privileged information such as people segments. We expect that the setup, methods, and results in this paper will be useful for further studies of representation bias in computer vision.

**Acknowledgements** This research was supported partially by a Google Faculty Award, DARPA (HR0011-18-9-0019), and gift funding from SAP Research and Leidos Inc. We also acknowledge fruitful discussions with members of the Human-Machine Intelligence group through the Institute for the Humanities and Global Cultures at the University of Virginia.

## References


[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *Conference on Fairness, Accountability and Transparency*, 2017.

[2] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *Conference on Fairness, Accountability and Transparency*, 2017.

[3] 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*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[4] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. *European Conference on Computer Vision (ECCV)*, 2018.

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

[6] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[7] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, 2015.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[9] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3323–3331, 2016.

[10] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1929–1938, 2018.

[11] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM, 2015.

[12] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[13] Tae-hoon Kim, Dongmin Kang, Kari Pulli, and Jonghyun Choi. Training with the invisibles: Obfuscating images to share safely for learning visual recognition models. *arXiv preprint arXiv:1901.00098*, 2019.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[15] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4069–4079, 2017.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[17] Ping Liu, Joey Tianyi Zhou, Ivor Wai-Hung Tsang, Zibo Meng, Shizhong Han, and Yan Tong. Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In *European Conference on Computer Vision*, pages 151–166. Springer, 2014.

[18] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–29, 2017.

[19] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2080–2089, 2018.

[20] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*, pages 2930–2939, 2016.

[21] Sebastian Palacio, Joachim Folz, Jörn Hees, Federico Raue, Damian Borth, and Andreas Dengel. What do deep networks like to see? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3108–3117, 2018.

[22] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer, 2012.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[24] Karen Ross and Cynthia Carter. Women and news: A long and winding road. *Media, Culture & Society*, 33(8):1148–1165, 2011.

[25] Hee Jung Ryu, Margaret Mitchell, and Hartwig Adam. Improving smiling detection with race and gender diversity. *arXiv*, 2017.

[26] Jure Sokolic, Qiang Qiu, Miguel RD Rodrigues, and Guillermo Sapiro. Learning to succeed while teaching to fail: Privacy in closed machine learning systems. *arXiv preprint arXiv:1705.08197*, 2017.

[27] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *arXiv preprint arXiv:1711.11443*, 2017.

[28] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

[29] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, pages 37–55. Springer, 2017.

[30] A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE Computer Society, 2011.

[31] Maneesh Upmanyu, Anoop M Namboodiri, Kannan Srinathan, and CV Jawahar. Efficient privacy preserving video surveillance. In *2009 IEEE 12th international conference on computer vision*, pages 1639–1646. IEEE, 2009.

[32] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

[33] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[34] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–596, 2017.

[35] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2925–2934, 2017.

[36] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[38] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *Proceedings of AIES*, 2018.

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

[40] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.