Representation Learning with Statistical Independence to Mitigate Bias

Ehsan Adeli[1,2*], Qingyu Zhao[1*], Adolf Pfefferbaum[1,3], Edith V. Sullivan[1],
Li Fei-Fei[2], Juan Carlos Niebles[2], Kilian M. Pohl[1,3]
[1]Department of Psychiatry and Behavioral Sciences, Stanford University, CA 94305
[2]Department of Computer Science, Stanford University, CA 94305
[3]Center for Biomedical Sciences, SRI International, Menlo Park, CA 94205
{eadeli, qingyuz, dolfp, edie, feifeili, jniebles, kilian.pohl}@stanford.edu

Abstract

Presence of bias (in datasets or tasks) is inarguably one of the most critical challenges in machine learning applications that has alluded to pivotal debates in recent years. Such challenges range from spurious associations between variables in medical studies to the bias of race in gender or face recognition systems. Controlling for all types of biases in the dataset curation stage is cumbersome and sometimes impossible. The alternative is to use the available data and build models incorporating fair representation learning. In this paper, we propose such a model based on adversarial training with two competing objectives to learn features that have (1) maximum discriminative power with respect to the task and (2) minimal statistical mean dependence with the protected (bias) variable(s). Our approach does so by incorporating a new adversarial loss function that encourages a vanished correlation between the bias and the learned features. We apply our method to synthetic data, medical images (containing task bias), and a dataset for gender classification (containing dataset bias). Our results show that the learned features by our method not only result in superior prediction performance but also are unbiased. The code is available at https://github.com/QingyuZhao/BR-Net/.

1. Introduction

A central challenge in practically all machine learning applications is how to identify and mitigate the effects of the bias present in the study. Bias can be defined as one or a set of extraneous protected variables that distort the relationship between the input (independent) and output (dependent) variables and hence lead to erroneous conclusions [39]. In a variety of applications ranging from disease prediction to face recognition, machine learning models are built to predict labels from images. Variables such as age, sex, and race may influence the training if the labels distribution is skewed with respect to them. Hence, the model may learn bias effects instead of actual discriminative cues.

The two most prevalent types of biases are dataset bias [44, 28] and task bias [31, 26]. *Dataset bias* is often introduced due to the lack of enough data points spanning the whole spectrum of variations with respect to one or a set of protected variables (*i.e.*, variables that define the bias). For example, a model that predicts gender from face images may have different recognition capabilities for different races with uneven sizes of training samples [10]. *Task bias*, on the other hand, is introduced by the intrinsic dependency between protected variables and the task. For instance, in neuroimaging applications, demographic variables such as gender [18] or age [16] are crucial protected variables; *i.e.*, they affect both the input (*e.g.*, neuroimages) and output (*e.g.*, diagnosis) of a prediction model so they likely introduce a distorted association. Both bias types pose serious challenges to learning algorithms.

With the rapid development of deep learning methods, Convolutional Neural Networks (CNN) are emerging as eminent ways of extracting representations (features) from imaging data. However, like other machine learning methods, CNNs are prone to capturing any bias present in the task or dataset when not properly controlled. Recent work has focused on methods for understanding causal effects of bias on databases [44, 27] or learning fair models [32, 50, 28, 44] with de-biased representations based on the developments in invariant feature learning [20, 6] and domain adversarial learning [19, 43]. These methods have shown great potential when the protected variables are dichotomous or categorical. However, their applications to handling task bias and continuous protected variables are still under-explored.

In this paper, we propose a representation learning scheme that learns features predictive of class labels with minimal bias to any generic type of protected variables. Our method is inspired by the domain-adversarial training approaches [20] with controllable invariance [55] within the

*Equal Contribution

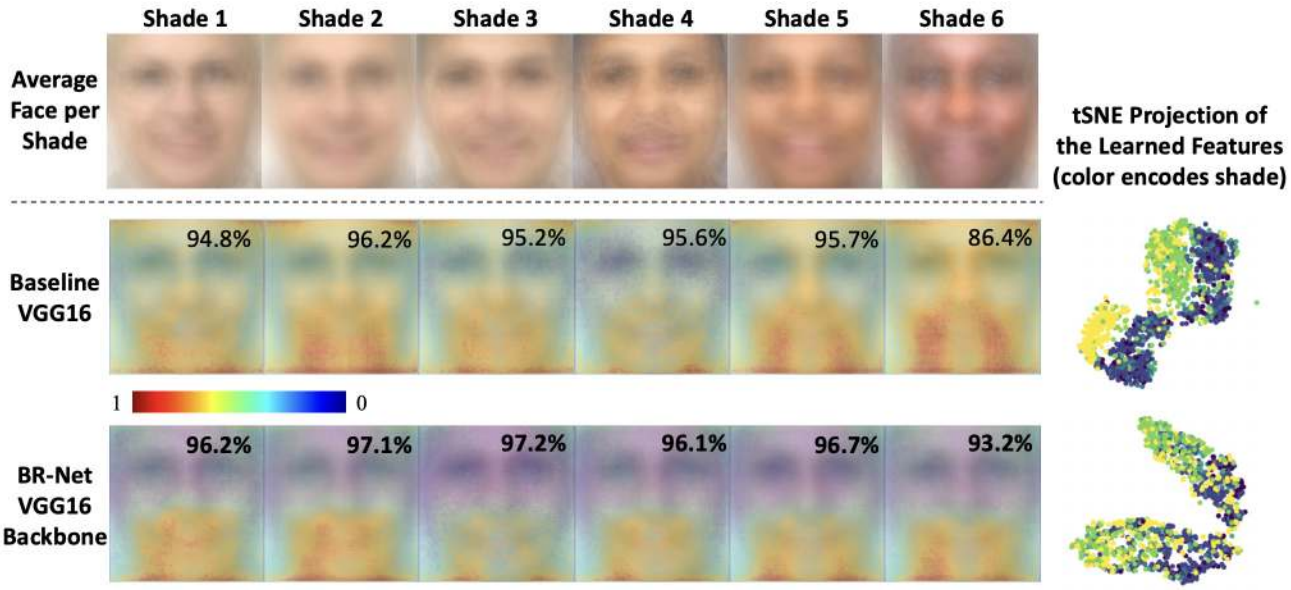


Figure 1: Average face images for each shade category (1^{st} row), average saliency map of the trained baseline (2^{nd} row), and BR-Net (3^{rd} row) color-coded with the normalized saliency for each pixel. BR-Net results in more stable patterns across all 6 shades. The last column shows the tSNE projection of the learned representations by each method. Our method results in a better representation space invariant to the bias variable (shade) while the baseline shows a pattern influenced by the bias. Average accuracy of per-shade gender classification over 5 runs of 5-fold cross-validation (pre-trained on ImageNet, fine-tuned on GS-PPB) is shown on each average map. BR-Net not only obtains better accuracy for the darker shade but also *regularizes* the model to improve overall per-category accuracy.

context of generative adversarial networks (GANs) [22]. We introduce an adversarial loss function based on the Pearson correlation between true and predicted values of a protected variable. Unlike prior methods, this strategy can handle protected variables that are continuous or ordinal. We theoretically show that the adversarial minimization of the linear correlation can remove non-linear association between the learned representations and protected variables, thus achieving *statistical mean independence*. Further, our strategy improves over the commonly used cross-entropy or mean-squared error (MSE) loss that is often *ill-posed* when optimized adversarially. Our method, denoted by Bias-Resilient Neural Network (BR-Net), uses architectures similar to the prior adversarial invariant feature learning works, but injects resilience towards the bias during training by taking an statistical independence approach to produce bias-invariant features at the presence of dataset and task biases. BR-Net is novel compared to the prior fair representation learning methods as (1) it can deal with *continuous* and *ordinal* protected variables and (2) is based on a theoretical proof of mean independence within the adversarial training context.

We evaluate BR-Net on two datasets that allow us to highlight different aspects of the method in comparison with a wide range of baselines. First, we test it on a *medical imaging application*, *i.e.*, distinguishing T1-weighted Magnetic Resonance Images (MRIs) of patients with the human immunodeficiency virus (HIV) from those of healthy subjects. As documented by the HIV literature, HIV accelerates the aging process of the brain [13], thereby introducing a task bias with respect to age (a continuous variable). In other words, if a predictor is trained not considering age as a protected variable (or confounder as referred to in medical studies), the predictor may actually learn the brain aging patterns rather than actual HIV markers. Then, we evaluate BR-Net for *gender classification* using the Gender Shades Pilot Parliaments Benchmark (GS-PPB) dataset [10]. We use different backbones pre-trained on ImageNet [15] in BR-Net and fine-tune them for our specific task, *i.e.*, gender prediction from face images. We show that prediction accuracy of the vanilla model is dependent on the subject's skin color (quantified by the 'shade' variable, an ordinal variable), which is not the case for BR-Net. Our comparison with several baselines and prior state-of-the-art shows that BR-Net not only learns features impartial to race (verified by feature embedding visualizations) but also results in higher accuracy (Fig. 1).

2. Related Work

Fairness in Machine Learning: In recent years, developing fair machine learning models have been the center of many discussions [33, 23, 3] including the media [29, 37]. It is often argued that human or society biases are replicated in the training datasets and hence can be seen in learned

models [4]. Recent effort in solving this problem focused on building fairer datasets [56, 11, 44]. However, this approach is not always practical for large-scale datasets or for applications, where data is relatively scarce and expensive to generate (*e.g.*, medical applications). Other works learn fair representations leveraging the existing data [58, 14, 53] by identifying features that are only predictive of the actual outputs, *i.e.*, impartial to the protected variable. But they come short when the protected variables are continuous.

Domain-Adversarial Training: [20] proposed for the first time to use adversarial training for domain adaptation tasks by using the learned features to predict the domain label (a binary variable; source or target). Several other works built on top of the same idea and explored different loss functions [9], domain discriminator settings [51, 17, 8], or cycle-consistency [25]. The focus of all these works was to close the domain gap, which is often encoded as a binary variable. To learn general-purpose bias-resilient models, we need new theoretical insight into the methods that can learn features invariant to all types of protected variables.

Invariant Representation Learning: There have been different attempts in the literature for learning representations that are invariant to specific factors in the data. For instance, [58] took an information obfuscation approach to obfuscate membership in the protected group of data during training, and [6, 40] introduced regularization-based methods. Recently, [55, 2, 59, 19, 12] proposed to use domain-adversarial training strategies for invariant feature learning. Some works [43, 52] used adversarial techniques based on similar loss functions as in domain adaptation to predict the exact values of the protected variables. For instance, [52] used a binary cross-entropy for removing effect of 'gender' and [43] used linear and kernelized least-square predictors as the adversarial component. Several methods based on optimizing equalized odds [36], entropy [48, 42] and mutual-information [47, 7, 38] were also widely used for fair representation learning. However, these methods are intractable for continuous or ordinal protected variables.

3. Bias-Resilient Neural Network (BR-Net)

Suppose we have an M-class classification problem, for which we have N pairs of training images and their corresponding target label(s): $\{(\mathbf{X}_i, \mathbf{y}_i)\}_{i=1}^N$. Assuming a set of k protected variables, denoted by a vector $\mathbf{b} \in \mathbb{R}^k$, to train a model for classifying each image while being impartial to the protected variables, we propose an *end-to-end* architecture (Fig. 2) similar to domain-adversarial training approaches [20]. Given the input image $\mathbf{X}$, the representation learning ($\mathbb{FE}$) module extracts a feature vector $\mathbf{F}$, on top of which a Classifier ($\mathbb{C}$) predicts the class label $\mathbf{y}$. Now, to guarantee that these features are not biased to $\mathbf{b}$, we build a network (denoted by $\mathbb{BP}$) with a new loss function that checks the

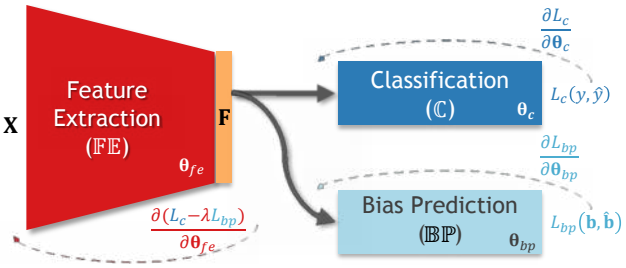
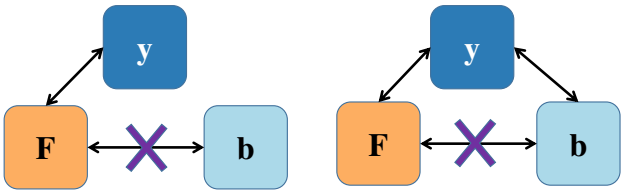


Figure 2: BR-Net architecture: $\mathbb{FE}$ learns features, $\mathbf{F}$, that successfully classify ($\mathbb{C}$) the input while being invariant (statistically independent) to the protected variables, $\mathbf{b}$, using $\mathbb{BP}$ and the adversarial loss, $-\lambda L_{bp}$ (based on correlation coefficient). Forward arrows show forward paths while the backward dashed ones indicate back-propagation with the respective gradient (∂) values.



(a) Dataset Bias (b) Task Bias

Figure 3: BR-Net can remove direct dependency between $\mathbf{F}$ and $\mathbf{b}$ for both dataset or task bias.

statistical mean dependence of the protected variables to $\mathbf{F}$. Back-propagating this loss to $\mathbb{FE}$ adversarially results in features that minimize the classification loss while having the least statistical dependence on the protected variables.

Each network has its underlying trainable parameters, defined as θ_{fe} for $\mathbb{FE}$, θ_c for $\mathbb{C}$, and θ_{bp} for $\mathbb{BP}$. If the predicted probability that subject i belongs to class m is defined by $\hat{y}_{im} = \mathbb{C}(\mathbb{FE}(\mathbf{X}_i; \theta_{fe}); \theta_c)$, the classification loss can be characterized by a cross-entropy:

$$L_c(\mathbf{X}, \mathbf{y}; \theta_{fe}, \theta_c) = -\sum_{i=1}^{N}\sum_{m=1}^{M} y_{im} \log(\hat{y}_{im}). \quad (1)$$

Similarly, with $\hat{\mathbf{b}}_i = \mathbb{BP}(\mathbb{FE}(\mathbf{X}_i; \theta_{fe}); \theta_{bp})$, we can define the adversarial component of the loss function. Standard methods for designing this loss function suggest to use a cross-entropy for binary/categorical variables (*e.g.*, in [20, 55, 12]) or an ℓ_2 MSE loss for continuous variables ([43]). These loss functions solely aim to maximize prediction error of $\mathbf{b}$ in the adversarial training but cannot remove statistical dependence. For example, the maximization of MSE between $\hat{\mathbf{b}}$ and $\mathbf{b}$ is an ill-posed (unbounded) objective and can be trivially achieved by uniformly shifting the magnitude of $\hat{\mathbf{b}}$, which does not remove any correlation with respect to $\mathbf{b}$. To avoid this issue, we define the surrogate loss for predicting the protected variables while quantifying the

statistical dependence with respect to $\mathbf{b}$ based on the squared Pearson correlation $\text{corr}^2(\cdot,\cdot)$:

$$L_{bp}(\mathbf{X},\mathbf{b};\theta_{fe},\theta_{bp}) = -\sum_{\kappa=1}^{k}\text{corr}^2(\mathbf{b}_\kappa,\hat{\mathbf{b}}_\kappa), \quad (2)$$

where $\mathbf{b}_\kappa$ defines the vector of κ^{th} protected variable across all N training inputs. Through adversarial training, we aim to remove statistical dependence by encouraging a zero correlation between $\mathbf{b}_\kappa$ and $\hat{\mathbf{b}}_\kappa$. Note, $\mathbb{BP}$ deems to maximize squared correlation and $\mathbb{FE}$ minimizes for it. Since corr^2 is bounded in the range [0, 1], both minimization and maximization schemes are feasible. Hence, the overall objective of the network is then defined as

$$\min_{\theta_{fe},\theta_c}\max_{\theta_{bp}} L_c(\mathbf{X},\mathbf{y};\theta_{fe},\theta_c) - \lambda L_{bp}(\mathbf{X},\mathbf{b};\theta_{fe},\theta_{bp}). \quad (3)$$

where hyperparameter λ controls the trade-off between the two objectives. This scheme is similar to GAN [22] and domain-adversarial training [20, 55], in which a min-max game is defined between two networks. In our case, $\mathbb{FE}$ extracts features that minimize the classification criterion, while 'fooling' $\mathbb{BP}$ (*i.e.*, making $\mathbb{BP}$ incapable of predicting the protected variables). Hence, the saddle point for this objective is obtained when the parameters θ_{fe} minimize the classification loss while maximizing the loss of $\mathbb{BP}$. Similar to the training of GANs, in each iteration, we first back-propagate the L_c loss to update θ_{fe} and θ_c. With θ_{fe} fixed, we then minimize the L_{bp} loss to update θ_{bp}. Finally, with θ_{bp} fixed, we maximize the L_{bp} loss to update θ_{fe}. In this study, L_{bp} depends on the correlation operation, which is a population-based operation, as opposed to individual-level error metrics such as cross-entropy or MSE losses. Therefore, we calculate the correlations over each training batch as a batch-level operation.

3.1. Non-linear Statistical Independence Guarantee

In general, a zero-correlation or a zero-covariance only quantifies linear independence between univariate variables but cannot infer non-linear relationships in high dimension. However, we now theoretically show that, under certain assumptions on the adversarial training of $\mathbb{BP}$, a zero-covariance would guarantee the *mean independence* [54] between protected variables and the high dimensional features, a much stronger type of statistical independence than the linear one.

A random variable $\mathcal{B}$ is said to be *mean independent* of $\mathcal{F}$ if and only if $E[\mathcal{B}|\mathcal{F}=\xi] = E[\mathcal{B}]$ for all ξ with non-zero probability, where $E[\cdot]$ defines the expected value. In other words, the expected value of $\mathcal{B}$ is neither linearly nor non-linearly dependent on $\mathcal{F}$, but the variance of $\mathcal{B}$ might. The following theorem then relates the mean independence between features $\mathcal{F}$ and the protected variables $\mathcal{B}$ to the zero-covariance between $\mathcal{B}$ and the $\mathbb{BP}$ prediction, $\hat{\mathcal{B}}$.

Property 1: $\mathcal{B}$ is mean independent of $\hat{\mathcal{B}} \Rightarrow Cov(\mathcal{B},\hat{\mathcal{B}}) = 0$.
Property 2: $\mathcal{B},\mathcal{F}$ are mean independent $\Rightarrow$ $\mathcal{B}$ is mean independent of $\hat{\mathcal{B}} = \phi(\mathcal{F})$ for any mapping function ϕ.

Theorem 1. *Given random variables $\mathcal{F},\mathcal{B},\hat{\mathcal{B}}$ with finite second moment, $\mathcal{B}$ is mean independent of $\mathcal{F}$ $\Leftrightarrow$ for any arbitrary mapping ϕ, s.t. $\hat{\mathcal{B}} = \phi(\mathcal{F})$, $cov(\mathcal{B},\hat{\mathcal{B}}) = 0$*

Proof. The forward direction $\Rightarrow$ follows directly through *Property* 1 and 2. We focus the proof on the reverse direction. Now, construct a mapping function $\hat{\mathcal{B}} = \phi(\mathcal{F}) = E[\mathcal{B}|\mathcal{F}]$, *i.e.*, $\phi(\xi) = E[\mathcal{B}|\mathcal{F}=\xi]$, then $Cov(\mathcal{B},\hat{\mathcal{B}}) = 0$ implies

$$E\big[\mathcal{B}E[\mathcal{B}|\mathcal{F}]\big] = E[\mathcal{B}]E\big[E[\mathcal{B}|\mathcal{F}]\big]. \quad (4)$$

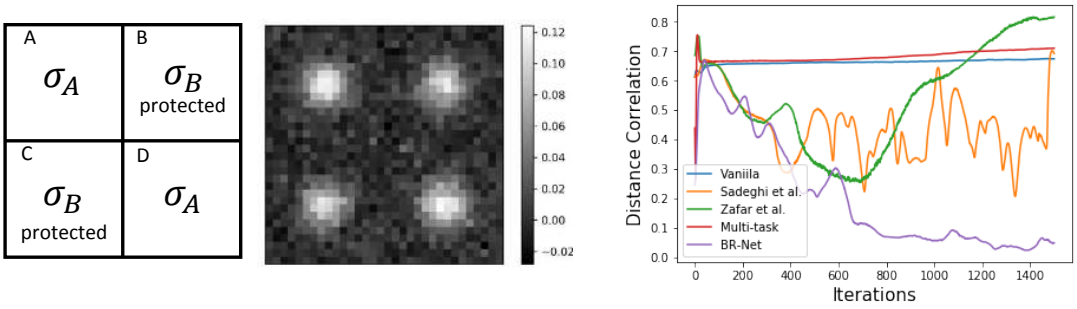
Due to the self-adjointness of the mapping $\mathcal{B}\mapsto E[\mathcal{B}|\mathcal{F}]$, the left hand side of Eq. (4) reads $E\big[\mathcal{B}E[\mathcal{B}|\mathcal{F}]\big] = E\big[\,(E[\mathcal{B}|\mathcal{F}])^2\,\big] = E[\hat{\mathcal{B}}^2]$. By the law of total expectation $E\big[E[\mathcal{B}|\mathcal{F}]\big] = E[\mathcal{B}]$, the right hand side of Eq. (4) becomes $E[\hat{\mathcal{B}}]^2$. By Jensen's (in)equality, $E[\hat{\mathcal{B}}^2] = E[\hat{\mathcal{B}}]^2$ holds iff $\hat{\mathcal{B}}$ is a constant, *i.e.*, $\mathcal{B}$ is mean independent of $\mathcal{F}$. □

Remark. *In practice, we normalize the covariance by standard deviations of variables for optimization stability. In the unlikely singular case that $\mathbb{BP}$ outputs a constant, we add a small perturbation in computing the standard deviation.*

This theorem echoes the validity of our adversarial training strategy: $\mathbb{FE}$ encourages a zero-correlation between $\mathbf{b}_\kappa$ and $\hat{\mathbf{b}}_\kappa$, which enforces $\mathbf{b}_\kappa$ to be mean independent of $\mathbf{F}$ (one cannot infer the expected value of $\mathbf{b}_\kappa$ from $\mathbf{F}$). In turn, assuming $\mathbb{BP}$ has the capacity to approximate any arbitrary mapping function, the mean independence between features and bias would correspond to a zero-correlation between $\mathbf{b}_\kappa$ and $\hat{\mathbf{b}}_\kappa$, otherwise $\mathbb{BP}$ would adversarially optimize for a mapping function that increases the correlation.

Moreover, **Theorem 1** induces that when $\mathbf{b}_\kappa$ is mean independent of $\mathbf{F}$, $\mathbf{b}_\kappa$ is also mean independent of $\mathbf{y}$ for any arbitrary classifier $\mathbb{C}$, indicating that the prediction is guaranteed to be unbiased. When $\mathbb{C}$ is a binary classifier and $\mathbf{y}\sim\text{Ber}(q)$, we have $p(\mathbf{y}=1|\mathbf{b}_\kappa) = E[\mathbf{y}|\mathbf{b}_\kappa] = E[\mathbf{y}] = p(\mathbf{y}=1) = q$; that is, $\mathbf{y}$ and $\mathbf{b}_\kappa$ are fully independent.

As mentioned, when $\mathbf{b}$ characterizes dataset bias (Fig. 2a), there is no intrinsic link between the protected variable and the task label (*e.g.*, in gender recognition, probability of being a female is not dependent on race), and the bias is introduced due to the data having a skewed distribution with respect to the protected variable. In this situation, we should train $\mathbb{BP}$ on the entire dataset to remove dependency between $\mathbf{F}$ and $\mathbf{b}$. On the other hand, when $\mathbf{b}$ is a task bias (Fig. 2b), it will have an intrinsic dependency with the task label (*e.g.*, in disease classification, the disease group has a different age range than the control group), such that the task label $\mathbf{y}$ could potentially become a moderator [5] that affects the strength of dependency between the features and protected variables.



(a) Synthetic data image synthesis.

(b) Distance correlation w.r.t. σ_B

Figure 4: Formation of synthetic dataset (a) and comparison of results for different methods (b).

In this situation, the goal of fair representation learning is to remove the direct statistical dependency between $\mathbf{F}$ and $\mathbf{b}$ while tolerating the indirect association induced by the task. Therefore, our adversarial training aims to ensure mean independence between $\mathbf{F}$ and $\mathbf{b}$ conditioned on the task label $E[\mathbf{F}|\mathbf{b},\mathbf{y}] = E[\mathbf{F}|\mathbf{y}]$, $E[\mathbf{b}|\mathbf{F},\mathbf{y}] = E[\mathbf{b}|\mathbf{y}]$. In practice, we train the adversarial loss within one or each of the M classes, depending on the specific task. This alleviates the 'competing equilibrium' issue in common fair machine learning methods [55], where the aim is to achieve full independence w.r.t $\mathbf{b}$ while accurately predict $\mathbf{y}$, an impossible task.

4. Experiments

We evaluate our method on two different scenarios. We compare BR-Net with several baseline approaches, and evaluate how our learned representations are invariant to the protected variables.

Baseline Methods. In line with the implementation of our approach, the baselines for all three experiments are 1) Vanilla: a vanilla CNN with an architecture exactly the same as BR-net without the bias prediction sub-network and hence the adversarial loss; and 2) Multi-task: a single $\mathbb{FE}$ followed by two separate predictors for predicting $\mathbf{b}_\kappa$ and $\mathbf{y}$, respectively [34]. The third type of approaches used for comparison are other unbiased representation learning methods. **Note that most existing works for "fair deep learning" are only designed for binary or categorical bias variables**. Therefore, in the synthetic and brain MRI experiments where the protected variable is continuous, we compare with two applicable scenarios originally proposed in the logistic regression setting: 1) [43] uses the MSE between the predicted and true bias as the adversarial loss; 2) [57] aims to minimize the magnitude of correlation between $\mathbf{b}_\kappa$ and the logit of $\mathbf{y}$, which in our case is achieved by adding the correlation magnitude to the loss function. For the Gender Shades PPB experiment, the protected variable is categorical. We then further compare with [30], which uses conditional entropy as the adversarial loss to minimize the mutual information between bias and features. Note, entropy-based [48, 42] and mutual-information-based methods [47, 7, 38] are widely used in fair representation learning to handle discrete bias.

Metrics for Accuracy and Statistical Independence. For the MRI and GS-PPB experiments, we measure prediction accuracy of each method by recording the balanced accuracy (bAcc), F_1-score, and AUC from a 5-fold cross-validation. In addition, we track the statistical dependency between the protected variable and features during the training process by applying the model to the entire dataset. We then compute the squared distance correlation ($dcor^2$) [49] and mutual information (MI) between the learned features at each iteration and the ground-truth protected variable. Note that the computation of $dcor^2$ and MI does not involve the bias predictor ($\mathbb{BP}$), thereby enabling a unified comparison between adversarial methods and the non-adversarial ones. Unlike Pearson correlation, $dcor^2 = 0$ or MI$= 0$ imply full statistical independence with respect to the features in the high dimensional space. Lastly, the discrete protected variable in GS-PPB experiment allows us to record another independence metric called the Equality of Opportunity (EO). EO measures the average gap in true positive rates w.r.t. different values of the protected variable.

4.1. Synthetic Experiments

We generate a synthetic dataset comprised of two groups of data, each containing 512 images of resolution 32×32 pixels. Each image is generated by 4 Gaussians (see Fig. 4a), the magnitude of which is controlled by σ_A and σ_B. For each image from Group 1, we sample σ_A and σ_B from a uniform distribution $\mathcal{U}(1,4)$ while we generate images of Group 2 with stronger intensities by sampling from $\mathcal{U}(3,6)$. Gaussian noise is added to the images with standard deviation 0.01. Now we assume the difference in σ_A between the two groups is associated with the true discriminative cues that should be learned by a classifier, whereas σ_B is a protected variable.

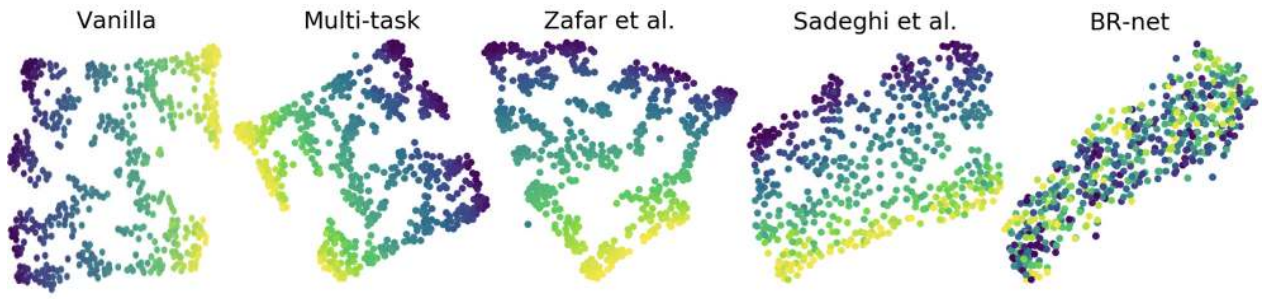


Figure 5: tSNE projection of the learned features for different methods. Color indicates the value of σ_B.

In other words, an unbiased model should predict the group label purely based on the two diagonal Gaussians and not dependent on the two off-diagonal ones. To show that the BR-Net can result in such models by controlling for σ_B, we train it on the whole dataset of 1,024 images with binary labels and σ_B values.

For simplicity, we construct $\mathbb{FE}$ with 3 stacks of 2×2 convolution/ReLU/max-pooling layers to produce 32 features. Both the $\mathbb{BP}$ and $\mathbb{C}$ networks have one hidden layer of dimension 16 with $tanh$ as the non-linear activation function. After training, the vanilla and multi-task models achieve close to 95% training accuracy, and the other 3 methods close to 90%. Note that the theoretically maximum training accuracy is 90% due to the overlapping sampling range of σ_A between the two groups, indicating that the vanilla and multi-task models additionally rely on the protected variable σ_B for predicting the group label, an undesired behavior. Further, Fig. 4b shows that our method can optimally remove the statistical association w.r.t. σ_B as $dcor^2$ drops dramatically with training iterations. The MSE-based adversarial loss yields unstable $dcor^2$ measures, and [57] suboptimally removes the bias in the features (green curve Fig. 4b). Finally, the above results are further supported by the 2D t-SNE [35] projection of the learned features as shown in Fig. 5. BR-net results in a feature space with no apparent bias, whereas features derived by other methods form a clear correlation with σ_B. This confirms the unbiased representation learned by BR-Net.

4.2. HIV Diagnosis Based on MRIs

Neuroimaging studies increasingly rely on machine learning models to identify differences in brain images between cohorts. Our first task aims at classifying HIV patients *vs*. control subjects (CTRL) based on brain MRIs to help understanding the impact of HIV on the brain. The study cohort includes 223 CTRLs and 122 HIV patients who are seropositive for the HIV-infection with CD4 count > 100 $\frac{\text{cells}}{\mu L}$ (average: 303.0). Since the HIV subjects are significantly older in age than the CTRLs (CTRL: 45 ± 17, HIV: 51 ± 8.3, $p < .001$) in this study, normal aging becomes a potential task bias; prediction of diagnosis labels may be dependent on subjects' age instead of true HIV markers.

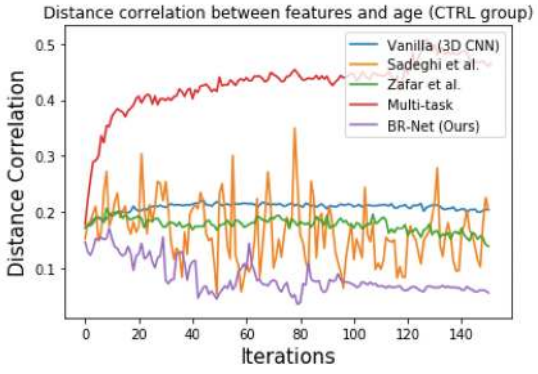


Figure 6: Statistical dependence between the learned features and age for the CTRL cohort in the HIV experiment, which is quantitatively measured by $dcor^2$.

The T1-weighted MRIs are all skull stripped, affinely registered to a common template, and resized into a $64 \times 64 \times 64$ volume. For each run of the 5-fold cross-validation, the training folds are augmented by random shifting (within one-voxel distance), rotation (within one degree) in all 3 directions, and left-right flipping based on the assumption that HIV infection affects the brain bilaterally [1]. The data augmentation results in a balanced training set of 1024 CTRLs and 1024 HIVs. As the flipping removes left-right orientation, the ConvNet is built on half of the 3D volume containing one hemisphere. The representation extractor $\mathbb{FE}$ has 4 stacks of $2 \times 2 \times 2$ 3D convolution/ReLu/batch-normalization/max-pooling layers yielding 4096 intermediate features. Both $\mathbb{BP}$ and $\mathbb{C}$ have one hidden layer of dimension 128 with $tanh$ as the activation function. As discussed, the task bias should be handled within individual groups rather the whole dataset. Motivated by recent medical studies [41, 1], we perform the adversarial training with respect to the protected variable of age only on the CTRLs because HIV subjects may exhibit irregular aging. Extension analysis of this conditional modeling applied to medical applications can be found at [60].

Table 1 shows the diagnosis prediction accuracy of BR-

Table 1: Classification accuracy of HIV diagnosis prediction and statistical dependency of learned features w.r.t. age. Best result in each column is typeset in bold and the second best is underlined.

Method	bAcc	F_1	AUC	dcor^2	MI
Vanilla	71.6	0.68	80.8	0.21	0.07
Multi-task	**74.2**	<u>0.66</u>	**82.5**	0.47	1.31
Sadeghi *et al*. [43]	64.8	0.58	75.2	0.22	0.06
Zafar *et al*. [57]	<u>73.2</u>	0.65	80.8	<u>0.15</u>	<u>0.04</u>
BR-Net (Ours)	**74.2**	**0.74**	<u>80.9</u>	**0.05**	**7e-4**

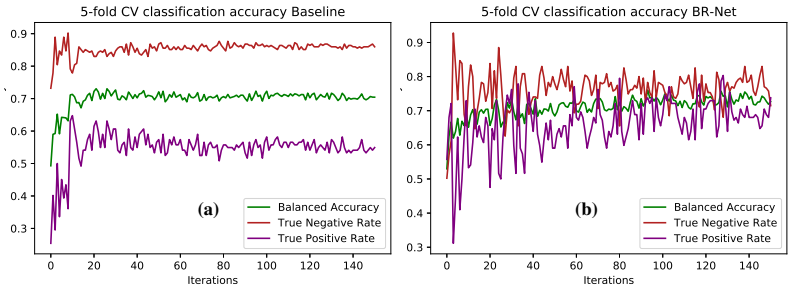


Figure 7: Accuracy, TNR, and TPR of the HIV experiment, as a function of the # of iterations for (a) 3D CNN baseline, (b) BR-Net. Our method is robust against the imbalanced age distribution between HIV and CTRL.

Net in comparison with baseline methods. BR-Net results in the most accurate prediction in terms of balanced accuracy (bAcc) and F_1-score, while it also learns the least biased features in terms of $dcor^2$ and MI. Although prior works [55] suggested that improving fairness of the model may reduce prediction accuracy due to the "competing equilibrium" between $\mathbb{C}$ and $\mathbb{BP}$, our results on this relatively small dataset indicate that naive classifiers may easily overfit to the aging (bias) effect and therefore result in lower accuracy in HIV classification. On the other hand, BR-Net alleviates the "competing equilibrium" issue in the task bias by pursuing conditional independence (CTRL) between features and age. While the multi-task model produces a higher AUC, it is also the most biased model as it simultaneously leverages both age and HIV cues for prediction. This result is also supported by Fig. 6, where the distance correlation for BR-Net decreases with the adversarial training and increases for Multi-task. The MSE-based adversarial loss [43] yields unstable $dcor^2$ measures potentially due to the ill-posed optimization of maximizing ℓ_2 distance. Moreover, minimizing the statistical association between the bias and predicted label $\mathbf{y}$ [57] does not necessarily lead to unbiased features (green curve Fig. 6). In addition, we record the true positive and true negative rate of BR-net for each training iteration. As shown in Fig. 7, the baseline tends to predict most subjects as CTRLs (high true negative rate). This is potentially caused by the CTRL group having a wider age distribution, so an age-dependent predictor would bias the prediction towards CTRL. When controlling for age, BR-Net reliably results in balanced true positive and true negative rates.

4.3. Gender Prediction Using the GS-PPB Dataset

The last experiment is on gender prediction from face images in the Gender Shades Pilot Parliaments Benchmark (GS-PPB) dataset [10]. This dataset contains 1,253 facial images of 561 female and 692 male subjects. The face shade is quantified by the Fitzpatrick six-point labeling system and is categorised from type 1 (lighter) to type 6 (darker). This quantization was used by dermatologists for skin classification and determining risk for skin cancer [10]. To ensure prediction is purely based on facial areas, we first perform face detection and crop the images [21]. To train our models on this dataset, we use backbones VGG16 [46] and ResNet50 [24] pre-trained on ImageNet [15]. We fine-tune each model on GS-PPB dataset to predict the gender of subjects based on their face images using fair 5-fold cross-validation. The ImageNet dataset for pre-training the models has fewer cases of humans with darker faces [56], and hence the resulting models have an underlying dataset bias to the shade.

BR-Net counts the variable 'shade' as an ordinal and categorical protected variable. As discussed earlier, besides the baseline models in the HIV experiment, we additionally compare with a fair representation learning method, [30], based on mutual information minimization. Note that this method is designed to handle discrete protected variables, therefore not applicable in previous experiments. We exclude [43] as the adversarial MSE-loss results in large oscillation in prediction results. Table 2 shows the prediction results across five runs of 5-fold cross-validation and the independence metrics derived by training on the entire dataset. Fig. 8 plots the accuracy for each individual 'shade' category. In terms of bAcc, BR-Net results in more accurate gender prediction than all baseline methods except that it it slightly worse than [57] with ResNet50 backbone. However, features learned by [57] are more biased towards skin shade. In most cases our method produces less biased features than [30], a method designed to explicitly optimize full statistical independence between variables. In practice, removing mean dependency by adversarial training is potentially a better surrogate for removing statistical dependency between high-dimensional features and bias.

BR-Net produces similar accuracy across all 'shade' categories. Prediction made by other methods, however, is more dependent on the protected variable by showing inconsistent recognition capabilities for different 'shade' categories and failing significantly on darker faces. This bias is confirmed by the t-SNE projection of the feature spaces (see Fig. 9) learned by the baseline methods; they all form clearer

Table 2: Average results over five runs of 5-fold cross-validation (accuracy and statistical independence metrics) on GS-PPB. Best results are typeset in bold and second best are underlined.

	VGG16 Backbone						ResNet50 Backbone					
Method	**bAcc (%)**	**F_1 (%)**	**AUC (%)**	$dcor^2$	**MI**	**EO%**	**bAcc (%)**	**F_1 (%)**	**AUC (%)**	$dcor^2$	**MI**	**EO%**
Vanilla	94.1±0.2	93.5±0.3	98.9±0.1	0.17	0.40	4.29	90.7±0.7	89.8±0.7	97.8±0.1	0.29	0.60	11.2
Kim *et al.* [30]	95.8±0.5	95.7±0.5	99.2±0.2	0.32	0.28	4.12	91.4±0.9	91.0±0.9	96.6±0.7	**0.18**	0.55	3.86
Zafar *et al.* [57]	94.3±0.4	93.7±0.5	99.0±0.1	0.19	0.43	4.11	**94.2±0.4**	**93.6±0.4**	**98.7±0.1**	0.29	0.60	4.68
Multi-Task	94.0±0.3	93.4±0.3	98.9±0.1	0.28	0.42	4.45	94.0±0.3	93.4±0.3	98.6±0.3	0.29	0.63	4.15
BR-Net	**96.3±0.6**	**96.0±0.7**	**99.4±0.2**	**0.12**	**0.13**	**2.02**	94.1±0.2	**93.6±0.2**	98.6±0.1	0.23	**0.49**	**2.87**

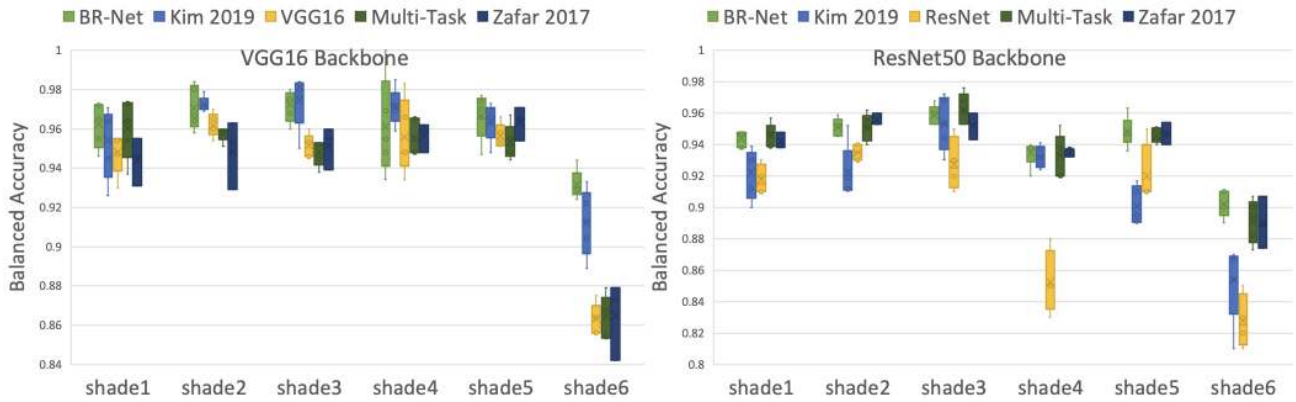


Figure 8: Accuracy of gender prediction from face images across all shades (1 to 6) of the GS-PPB dataset with two backbones, (left) VGG16 and (right) ResNet50. BR-Net consistently results in more accurate predictions in all 6 shade categories.

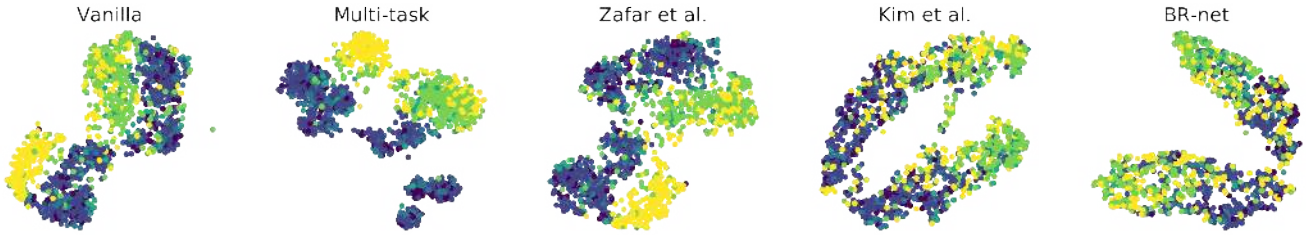


Figure 9: Learned representations by different methods. Color encodes the 6 categories of skin shade.

association with the bias variable than BR-Net. To gain more insight, we visualize the saliency maps derived for the baseline and BR-Net. For this purpose, we use a similar technique as in [45] to extract the pixels in the original image space highlighting the areas that are discriminative for the gender labels. Generating such saliency maps for all inputs, we visualize the average map for each individual 'shade' category (Fig. 1). The value on each pixel corresponds to the attention from the network to that pixel within the classification process. Compared to the baseline, BR-Net focuses more on specific face regions and results in more stable patterns across all 'shade' categories.

5. Conclusion

Machine learning models are acceding to everyday lives from policy making to crucial medical applications. Failure to account for the underlying bias in datasets and tasks can lead to spurious associations and erroneous decisions. We proposed a method based on adversarial training strategies by encouraging vanished correlation to learn features for the prediction task while being unbiased to the protected variables in the study. We evaluated our bias-resilient neural network (BR-Net) on synthetic, medical diagnosis, and gender classification datasets. In all experiments, BR-Net resulted in representations that were invariant to the protected variable while obtaining comparable (and sometime better) classification accuracy.

Acknowledgement. This work was supported in part by NIH Grants AA017347, AA010723, MH113406, and AA021697, and by Stanford HAI AWS Cloud Credit.

References

[1] Ehsan Adeli, Dongjin Kwon, Qingyu Zhao, Adolf Pfefferbaum, Natalie M Zahr, Edith V Sullivan, and Kilian M Pohl. Chained regularization for identifying brain patterns specific to HIV infection. *NeuroImage*, 183:425–437, 2018.

[2] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. *arXiv preprint arXiv:1904.12543*, 2019.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Neural Information Processing Systems Tutorial*, 2017.

[4] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[5] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[6] Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044*, pages 1733–1782, 2017.

[7] Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 614–623, 2019.

[8] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[9] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Pecognition*, pages 3722–3731, 2017.

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

[11] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

[12] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems*, pages 851–863, 2019.

[13] James H Cole, Jonathan Underwood, Matthan WA Caan, Davide De Francesco, Rosan A van Zoest, Robert Leech, Ferdinand WNM Wit, Peter Portegies, Gert J Geurtsen, Ben A Schmand, et al. Increased brain-predicted aging in treated HIV disease. *Neurology*, 88(14):1349–1357, 2017.

[14] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1436–1445, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[16] Beata Dobrowolska, Bernadeta J**k**edrzejkiewicz, Anna Pilewska-Kozak, Danuta Zarzycka, Barbara Ślusarska, Alina Deluga, Aneta Kościołek, and Alvisa Palese. Age discrimination in healthcare institutions perceived by seniors and students. *Nursing ethics*, 26(2):443–459, 2019.

[17] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

[18] Margrit Eichler, Anna Lisa Reisman, and Elaine Manace Borins. Gender bias in medical research. *Women & Therapy*, 12(4):61–70, 1992.

[19] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

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

[21] Adam Geitgey. Face recognition, https://pypi.org/project/face_recognition/, 2018.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[23] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the International Conference on Machine Learning*, 2018.

[24] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 5, pages 770–778, 2015.

[25] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[26] Fei Huang and Alexander Yates. Biased representation learning for domain adaptation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1313–1323. Association for Computational Linguistics, 2012.

[27] Aria Khademi and Vasant Honavar. Algorithmic bias in recidivism prediction: A causal perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13839–13840, Apr. 2020.

[28] 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.

[29] Dhruv Khullar. A.I. could worsen health disparities. *New York Times*, 2019.

[30] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[31] Nicol Turner Lee, Paul Resnick, and Genie Barton. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Center for Technology Innovation, Brookings. Tillgänglig online: https://www. brookings. edu/research/algorithmic-bias-detection-and-mitigation-bestpractices-and-policies-to-reduce-consumer-harms/# footnote-7 (2019-10-01)*, 2019.

[32] Yi Li and Nuno Vasconcelos. REPAIR: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019.

[33] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the International Conference on Machine Learning*, 2018.

[34] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5334–5343, 2017.

[35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[36] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[37] Clair Miller. Algorithms and bias. *New York Times*, 2015.

[38] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9084–9093. 2018.

[39] Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*, 5(2):79, 2012.

[40] Marc'aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[41] Anil Rao et al. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, 150:23–49, 2017.

[42] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. *CoRR*, abs/1904.05514, 2019.

[43] Bashir Sadeghi, Runyi Yu, and Vishnu Boddeti. On the global optima of kernelized adversarial representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7971–7979, 2019.

[44] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

[45] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[47] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89, pages 2164–2173. PMLR, 2019.

[48] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

[49] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[50] 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.

[51] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[52] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5310–5319, 2019.

[53] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *arXiv preprint arXiv:1912.07249*, 2019.

[54] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data (2nd ed.)*. The MIT Press, London, 2010.

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

[56] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. *Image-Net*, 2019.

[57] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry

Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970. PMLR, 2017.

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

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

[60] Qingyu Zhao*, Ehsan Adeli*, and Kilian M Pohl. Training confounder-free deep learning models for medical applications. *Nature Communications*, 2020 In Press.