

# Controllable Feature Whitening for Hyperparameter-Free Bias Mitigation

Yooshin Cho Hanbyel Cho Janghyeon Lee HyeongGwon Hong Jaesung Ahn Junmo Kim

Korea Advanced Institute of Science and Technology (KAIST)

{choys95, tlr14658, wkdgus9305, honggudrnjs, jaesung2, junmo.kim}@kaist.ac.kr

## Abstract

*As the use of artificial intelligence rapidly increases, the development of trustworthy artificial intelligence has become important. However, recent studies have shown that deep neural networks are susceptible to learn spurious correlations present in datasets. To improve the reliability, we propose a simple yet effective framework called controllable feature whitening. We quantify the linear correlation between the target and bias features by the covariance matrix, and eliminate it through the whitening module. Our results systemically demonstrate that removing the linear correlations between features fed into the last linear classifier significantly mitigates the bias, while avoiding the need to model intractable higher-order dependencies. A particular advantage of the proposed method is that it does not require regularization terms or adversarial learning, which often leads to unstable optimization in practice. Furthermore, we show that two fairness criteria, demographic parity and equalized odds, can be effectively handled by whitening with the re-weighted covariance matrix. Consequently, our method controls the trade-off between the utility and fairness of algorithms by adjusting the weighting coefficient. Finally, we validate that our method outperforms existing approaches on four benchmark datasets: Corrupted CIFAR-10, Biased FFHQ, WaterBirds, and Celeb-A.*

## 1. Introduction

Deep neural networks have shown impressive performance by capturing task-relevant statistical cues from well-curated training datasets [16, 56]. However, if training datasets are poorly curated, neural networks often rely on spurious cues that do not generalize well beyond the training distribution. Nevertheless, it is challenging to determine which statistical cues are beneficial for task performance; thus neural networks often fail when train dataset is highly biased (i.e., datasets in which a target attribute has a strong spurious correlation with a particular bias attribute). For example, if neural networks rely on spurious correlations to predict

the target attribute (e.g., recognizing objects by relying on backgrounds or textures), the generalization capability of the neural networks is severely reduced [46, 62]. Furthermore, previous studies have empirically demonstrated that neural networks tend to focus on *easier concepts* [1, 32], and over rely on such spurious correlations [36, 40].

To address the issues, several studies have been proposed [2, 11, 36, 40]. A common strategy is to enforce networks to learn representations that are independent to the specified bias attributes by incorporating fairness criteria (e.g., *demographic parity*, *equalized odds*, and *equal opportunity*) [2, 29, 45, 61]. Previous works have quantified the fairness criteria using statistical measures (e.g., mutual information, Hilbert Schmidt Independence Criterion, and Hirschfeld-Gebelein-Rényi coefficient) that represent the dependency between model predictions (or representations) and bias attributes. However, these measures are often analytically intractable or computationally expensive to estimate directly. Therefore, they employed neural networks to estimate the measures, and achieved fairness by adopting adversarial learning or regularization terms. However, it should be noted that adversarial learning can be easily unstable, and careful tuning of hyperparameters is required for regularization terms. Furthermore, it is difficult to evaluate whether the neural estimator precisely estimates the dependency during the min-max game.

To overcome the limitations, we propose a simple yet effective feature whitening based approach that is robust to hyperparameter tunings and facilitates stable training. As noted in prior works, linear independence can be satisfied with the whitening transform that multiplies the inverse square root of the covariance matrix [19, 20]. Although linear independence does not guarantee statistical independence, it ensures that one variable can not be estimated by a linear layer that takes other variables as inputs. This property allows our approach to function similarly to adversarial training, forcing the target representation to "forget" bias attributes in a linear regime. Furthermore, deep neural networks have empirically shown the capability to encode inputs into a representation that is linearly separable by a last linear classifier. Based on these insights, we estimate and

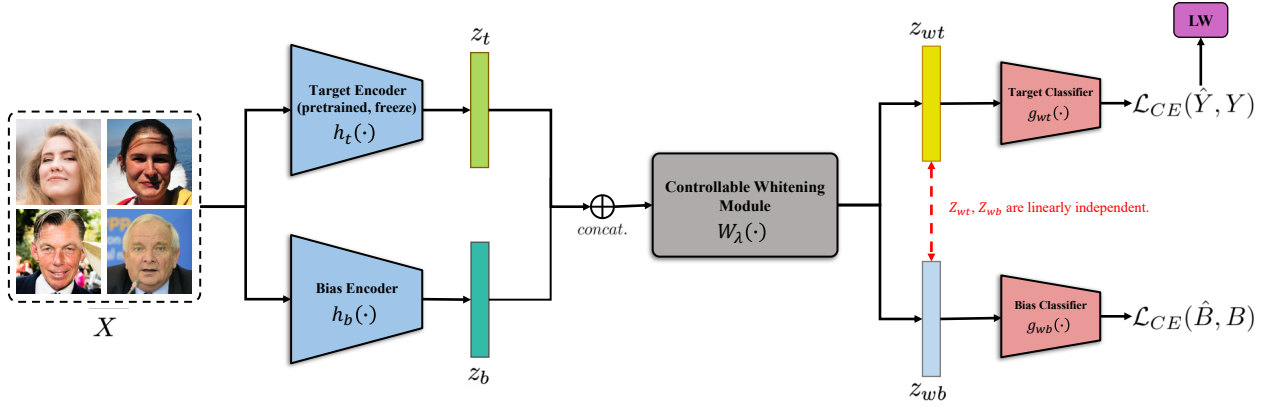


Figure 1. **Overview of proposed method.**  $X$  is the mini-batch of the images that sampled from the biased training dataset.  $h_t(\cdot)$  and  $h_b(\cdot)$  are target and bias encoder, respectively. To reduce the dependency, we remove the linear correlation between the target feature  $z_t$  and bias feature  $z_b$  using the controllable whitening module  $W_\lambda(\cdot)$ , that can handle *demographic parity* and *equalized odds* by controlling the coefficient  $\lambda$ . Subsequently, the whitened target feature  $z_{wt}$  and whitened bias feature  $z_{wb}$  are linearly independent, while  $z_{wt}$  is kept close to  $z_t$  by the coupled Newton-Schultz iteration that utilizes the degree of freedom in the  $\Sigma^{-1/2}$ . Then, we train linear classifiers,  $g_{wt}(\cdot)$  and  $g_{wb}(\cdot)$ , to predict the target attribute and bias attribute, respectively.  $LW$  refers to the loss weighting. To make learning stable, we freeze  $h_t(\cdot)$  which is pretrained on the same biased dataset.

eliminate the linear dependency between the target and bias features fed into the last linear classifiers, as illustrated in Figure 1. Notably, we demonstrate that significant improvements in fairness can be achieved by removing only linear dependency, without involving analytically intractable higher-order dependencies, when the whitening transform is appropriately applied.

In addition, we investigate the efficacy of the proposed method by evaluating *demographic parity*, which is one of the most widely used fairness criteria. *Demographic parity* requires the independence between the model predictions and the bias attribute. Experimental results confirm that our method effectively reduces *demographic parity* as the bias feature is trained to be linearly classified according to the bias attribute. However, it is well known that enforcement of strict *demographic parity* cripples the utility of the algorithm, particularly when the training dataset is highly biased, as it suppresses target-relevant information correlated with the bias attributes [15, 50]. By contrast, *equalized odds* means conditional independence between predictions and bias attributes given the target attributes; this does not conflict with learning target tasks regardless of the degree of the dataset bias [15].

To preserve task-relevant information while mitigating bias, we introduce a re-weighting strategy for covariance estimation. Specifically, we approximate a covariance matrix over the unbiased distribution by over-weighting the rare groups and under-weighting the predominant groups. In this paper, we denote the covariance matrix computed over the unbiased distribution (i.e., the target and bias at-

tributes are independent) as the unbiased covariance matrix. Since *equalized odds* and *demographic parity* become equivalent in an unbiased distribution, whitening with the unbiased covariance matrix naturally promotes *equalized odds*. Moreover, we empirically verify that whitening with the unbiased covariance matrix improves *equalized odds*, and prevents the task-relevant information loss. However, we also confirm that whitening with the purely unbiased covariance can lead to over-fitting, due to the sample diversity imbalance between groups which can not be mitigated by re-weighting.

To balance the trade-off between task-relevant information loss and over-fitting, we propose the Controllable Feature Whitening (CFW), which blends the unbiased and biased covariance matrices via a weighted arithmetic mean. By adjusting the weighting coefficient, we achieve a smoothly interpolated objective between *demographic parity* and *equalized odds*, enabling the model to mitigate inter-group performance disparities while simultaneously preserving the overall performance. Finally, we verify the efficacy of the proposed method by comparing the performance with existing methods on four benchmark datasets, Corrupted CIFAR-10 [34, 40], Biased FFHQ [25, 30], WaterBirds [49], and Celeb-A [38]. Empirically, we observe that setting the weighting coefficient to 0.25 consistently yields strong performance across datasets, suggesting that our method can be considered hyperparameter-free in practice. Notably, we demonstrate that its effectiveness can be further enhanced when integrated with other existing methods which improve the quality of the representation.

## 2. Related Work

**Fairness with known biases.** Many approaches to improve fairness assume that bias attributes (e.g., gender, age, and race) are known during training. The most intuitive strategies are re-weighting and re-sampling [6, 7, 47], adjusting the training distribution to mitigate bias. However, re-weighting and re-sampling often lead networks to over-fitting due to the lack of the sample diversity of rare groups. To address the issues, data augmentation techniques such as BiasSwap [30] and FlowAug [9] generate counterfactual examples to enhance the sample diversity. However, data generation may adversely affect learning, depending on the quality of the generated data. Adversarial learning, another widely used approach, trains an auxiliary branch to predict bias attributes, forcing the target network to forget the bias information [29, 61]. Although adversarial learning based approaches have demonstrated strong performance, adversarial learning is inherently unstable and requires careful tuning of hyperparameters. In this paper, we apply a re-weighting strategy when estimating the unbiased covariance matrix, and demonstrate reduced over-fitting compared to standard loss re-weighting. Furthermore, we replace adversarial learning with feature whitening in a linear regime, and simplify the training by eliminating the need of additional hyperparameters.

**Fairness with unknown biases.** On the other hand, recent studies have focused on more challenging yet practical scenarios in which bias attributes are not provided during training. In such cases, models should infer bias attributes from datasets by leveraging properties of biased networks. Loss-based methods identify bias by detecting high-loss samples, assuming that they correspond to underrepresented groups [37, 40]. To amplify the loss difference between groups, the Generalized Cross Entropy (GCE) loss [60] is widely utilized, and a committee of biased classifiers [31] is introduced. However, these methods struggle to distinguish between truly biased and intrinsically difficult examples, making them sensitive to hyperparameter tuning and requiring a labeled validation set for effective calibration. Cluster-based approaches, such as Correct-n-Contrast [59], seek to discover hidden bias attributes by grouping samples based on feature similarities, although such clustering may not always align with real-world biases. While fairness techniques that do not require explicit bias information have been increasingly studied and have shown even better performance than previous studies using the bias label, the absence of bias information can lead to limited generalization. Therefore, in this paper, we focus on the scenarios in which bias labels are provided, ensuring a more controlled approach to fairness.

## 3. Preliminary

### 3.1. Whitening and Independence

Whitening is a popular normalization technique which is widely adopted in various areas, including efficient optimization, domain adaptation, GAN, style transfer, and representation learning [10, 13, 48, 51]. It transforms the input features to have a zero mean and unit variance and removes the linear correlation between channels. It can be expressed by the following equation:

$$\tilde{\mathbf{X}} = \Sigma^{-\frac{1}{2}} \cdot (\mathbf{X} - \boldsymbol{\mu} \cdot \mathbf{1}^\top), \quad (1)$$

where  $\mathbf{X} \in \mathbb{R}^{C \times N}$  denotes the input features,  $N$  denotes the number of inputs, and  $C$  is the dimension size of inputs.  $\Sigma = \frac{1}{N}(\mathbf{X} - \boldsymbol{\mu} \cdot \mathbf{1}^\top) \cdot (\mathbf{X} - \boldsymbol{\mu} \cdot \mathbf{1}^\top)^\top$  and  $\boldsymbol{\mu} = \frac{1}{N}\mathbf{X} \cdot \mathbf{1}$  are the covariance matrix and mean vector of the input, respectively. Since the inverse square root of matrix is not unique (as multiplying any unitary matrix generates a valid inverse square root), many studies have been proposed such as ZCA-whitening [4, 26], Cholesky decomposition [12], and Newton Schulz iterations [5]. We employ the coupled Newton-Schultz iterations [17, 18] which is known to be numerically stable and computationally efficient [55].

Although zero covariance does not imply statistical independence between variables, it does ensure linear independence, which means that one variable can not be defined as a linear combination of the others (i.e., variables linearly unlearn each other). Moreover, if the variables follow a Gaussian distribution, a zero covariance implies statistical independence. Previous studies have analyzed that infinite-width neural network can be approximated as a Gaussian process by using the Central Limit Theorem [23, 35]. Although this does not rigorously match our work, we empirically demonstrate that whitening improves the fairness.

### 3.2. Fairness Criterion

**Problem Setup.** Let  $(X, Y, B) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{B}$  denote the input data, target attribute, and bias attribute, respectively, sampled from the dataset  $\mathcal{D}$ .  $\hat{Y} \in \mathcal{Y}$  denotes the algorithm prediction. We say that dataset  $\mathcal{D}$  is biased towards  $B$ , when  $Y$  and  $B$  are not independent (i.e.,  $P(Y|\mathcal{D}) \cdot P(B|\mathcal{D}) \neq P(Y, B|\mathcal{D})$ ). If a network is trained on a highly biased dataset,  $B$  can be used as a shortcut to predict  $Y$ . Following [36, 40], we refer to data samples as *bias-aligned* if  $Y$  can be correctly predicted by relying on  $B$ , which are predominant in biased datasets. Conversely, we say data samples are *bias-conflicting*, if the  $Y$  can not be predicted by relying on  $B$ , which are rare in biased datasets. If predictions,  $\hat{Y}$ , are highly biased toward  $B$ , the network poorly performs on *bias-conflicting* samples. Therefore, the objective of this work is not only improving the overall performance, but also reducing the performance gap between the groups.

**Fairness Criteria.** Numerous fairness criteria have been proposed to measure the group fairness of algorithms, including *demographic parity* [28, 57], *predictive parity* [14], *equalized odds*, and *equal opportunity* [15]. In the paper, we focus on two representative criteria: *demographic parity* and *equalized odds*, which are widely adopted to regularize training.

First, *demographic parity* requires the difference of prediction probability between the bias groups to be zero. This can be expressed using the following equation:

$$P(\hat{Y} = y|B = b_1) = P(\hat{Y} = y|B = b_2), \quad (2)$$

for  $\forall(y, b_1, b_2) \in \mathcal{Y} \times \mathcal{B} \times \mathcal{B}$ . This condition implies that the target prediction  $\hat{Y}$  and bias  $B$  are independent. Many successful approaches have been proposed to reduce the statistical dependency between  $B$  and  $\hat{Y}$  (or  $B$  and the target features) to achieve fairness. However, because  $B$  and  $Y$  are highly correlated in a biased training dataset, removing the dependency between  $B$  and  $\hat{Y}$  over the biased dataset inevitably reduces the dependency between  $Y$  and  $\hat{Y}$  [15] as well. In this paper, we quantify the degree of violation for *demographic parity* as the following:

$$\Delta_{DP} = \frac{1}{N_Y} \sum_{y \in \mathcal{Y}} \max_{b_1, b_2} |P(\hat{Y} = y|B = b_1) - P(\hat{Y} = y|B = b_2)|, \quad (3)$$

where  $N_y$  is the number of the classes of  $Y$ .

On the other hand, *equalized odds* requires the true positive ratio and false positive ratio over different bias groups to be the same, which can be expressed as the following equation:

$$\begin{aligned} &P(\hat{Y} = y_1|B = b_1, Y = y_2) \\ &= P(\hat{Y} = y_1|B = b_2, Y = y_2), \end{aligned} \quad (4)$$

for  $\forall(y_1, y_2, b_1, b_2) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{B} \times \mathcal{B}$ . *Equalized odds* implies that  $\hat{Y}$  and  $B$  are conditionally independent given  $Y$ . Owing to the conditioning on  $Y$ , this criterion preserves the dependency between  $Y$  and  $\hat{Y}$ , mitigating the risk of reducing predictive performance. In particular, *demographic parity* and *equalized odds* are equivalent if and only if  $Y$  and  $B$  are independent (i.e., dataset is unbiased). We quantify the degree of the violation for *equalized odds* as the following:

$$\begin{aligned} \Delta_{EO} &= \frac{1}{N_Y} \sum_{y \in \mathcal{Y}} \max_{b_1, b_2} |P(\hat{Y} = y|B = b_1, Y = y) \\ &\quad - P(\hat{Y} = y|B = b_2, Y = y)|. \end{aligned} \quad (5)$$

## 4. Methodology

### 4.1. Training Debaised Classifier

In this section, we describe the proposed method in detail. The key component of our approach is the whitening module, which makes the features linearly independent between channels without requiring unstable adversarial learning or regularization terms. By leveraging this property, we just need to train the networks to predict the target and bias attributes using the different groups of the channels of the whitened features. As one group of whitened features is trained to linearly classify the bias attribute, the other group inherently becomes incapable of linearly encoding bias information. The overall framework of the proposed method is illustrated in Figure 1.

Specifically, we train the network  $f_t(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ , which takes  $X$  as the input and predicts the target attribute  $Y$ , over the biased dataset  $\mathcal{D}_b$  using the standard cross-entropy loss.  $f_t(\cdot)$  is composed of the encoder network  $h_t(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^M$  and the linear classifier  $g_t(\cdot) : \mathbb{R}^M \rightarrow \mathcal{Y}$ , where  $M$  is the dimension size of the extracted target feature  $z_t = h_t(X)$ . Since,  $Y$  and  $B$  are highly correlated on  $\mathcal{D}_b$ , the network  $f_t(\cdot)$  is likely to make predictions by relying on  $B$  (i.e.,  $f_t(\cdot)$  shows great performance on the *bias-aligned* samples, but poor performance on the *bias-conflicting* samples). We refer this biased network  $f_t(\cdot)$  as *Vanilla* network. We bring the biased target encoder network  $h_t(\cdot)$  from *Vanilla* network and do not update, because satisfying fairness constraints, especially *demographic parity*, is known to easily conflict with learning target tasks [15]. Moreover, previous works have observed that fine-tuning the last linear layer is sufficient to achieve fairness [33, 52].

To mitigate the over-reliance of the pretrained *Vanilla* network, we remove the bias information from the target feature  $z_t$  using the feature whitening. We extract the bias feature  $z_b = h_b(X)$  using another bias encoder network  $h_b(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^M$ , which is trained to predict  $B$ . Then, the Controllable Feature Whitening (CFW)  $W_\lambda(\cdot) : \mathbb{R}^{2M} \rightarrow \mathbb{R}^{2M}$  takes the concatenated feature  $z = [z_t; z_b]$  as the input, and performs whitening using the Eq 1. Detailed explanations of the whitening process will be provided in Section 4.2. Consequently, the whitened feature  $z_w = W(z)$  satisfies the orthogonality between all channel pairs, and we split the  $z_w$  into whitened target feature  $z_{wt}$  and whitened bias feature  $z_{wb}$ . Then, the whitened target feature  $z_{wt}$  and whitened bias feature  $z_{wb}$  are linearly independent, while  $z_{wt}$  is kept close to  $z_t$  by the coupled Newton-Schultz iteration that utilizes the degree of freedom in the  $\Sigma^{-1/2}$  in the whitening module. Owing to the linear independence,  $z_{wt}$  and  $z_{wb}$  can not be estimated by linear layers takes each other as the input.

Then, we train the linear target classifier  $g_{wt}(\cdot) : \mathbb{R}^M \rightarrow \mathcal{Y}$ , which takes  $z_{wt}$  as the input and predicts the  $Y$ . Simi-

larly, we train the linear bias classifier  $g_{wb}(\cdot) : \mathbb{R}^M \rightarrow \mathcal{B}$ , which takes  $z_{wb}$  as the input and predicts  $B$ . For evaluation, we use  $\hat{Y} = g_{wt}(z_{wt})$ . The objective is composed of two terms:  $\mathcal{L}_t(h_t, g_{wt}, X, Y) = \mathcal{L}_{CE}(g_{wt}(z_{wt}), Y)$  and  $\mathcal{L}_b(h_b, g_{wb}, X, B) = \mathcal{L}_{CE}(g_{wb}(z_{wb}), B)$ , where  $[z_{wt}; z_{wb}] = W([h_t(X); h_b(X)])$ , and  $\mathcal{L}_{CE}$  is the cross-entropy loss. Finally the objective function can be written as the follows:

$$\min_{g_{wt}} \mathcal{L}_t(h_t, g_{wt}, X, Y) + \min_{h_b, g_{wb}} \mathcal{L}_b(h_b, g_{wb}, X, B). \quad (6)$$

## 4.2. Covariance Estimation and Re-weighting

We improve fairness of the network by whitening with the biased covariance matrix which is estimated over the biased training dataset in Section 4.1. However, as we mentioned, removing the correlation between  $\hat{Y}$  and  $B$  can conflict with learning the target task when training dataset is highly biased. The problem is caused by the fact that  $Y$  and  $B$  are not independent in the training dataset. To avoid the problem, we propose the controllable covariance estimation using re-weighting. The re-weighting strategy is simple but effective method to mimic the statistics of unbiased dataset. We simply over-weight the rare groups (i.e., *bias-conflicting* samples), and under-weight the common groups (i.e., *bias-aligned* samples). The biased covariance matrix  $\Sigma_b$  and the unbiased covariance matrix  $\Sigma_u$  can be expressed as follows:

$$\Sigma_b = \sum_{y, b \in \mathcal{Y}, \mathcal{B}} P(y, b | \mathcal{D}_b) \cdot \mathbb{E}_{\mathbf{X}_c \sim \mathcal{D}_b^{y, b}} [\mathbf{X}_c \mathbf{X}_c^T], \quad (7)$$

$$\Sigma_u = \sum_{y, b \in \mathcal{Y}, \mathcal{B}} P(y, b | \mathcal{D}_u) \cdot \mathbb{E}_{\mathbf{X}_c \sim \mathcal{D}_b^{y, b}} [\mathbf{X}_c \mathbf{X}_c^T], \quad (8)$$

where  $\mathcal{D}_b$  and  $\mathcal{D}_u$  are the biased and unbiased distributions, respectively.  $\mathcal{D}_b^{y, b}$  is the subset of  $\mathcal{D}_b$  that contains samples with  $Y = y$  and  $B = b$ . Thus,  $P(y, b | \mathcal{D}_b)$  can be obtained from the training dataset statistics. To ensure independence, we set  $P(y, b | \mathcal{D}_u) = \frac{1}{N_Y \cdot N_B}$ , where  $N_Y$  and  $N_B$  are the number of the classes of  $Y$  and  $B$ , respectively. To obtain a more general expression of the covariance matrix for mixed distributions, we add a weight coefficient  $\lambda \in [0, 1]$  to compute the weighted arithmetic mean of the biased and unbiased covariance matrices as the following:

$$\Sigma_\lambda = \lambda \cdot \Sigma_u + (1 - \lambda) \cdot \Sigma_b. \quad (9)$$

Consequently, the proposed Controllable Feature Whitening (CFW) performs whitening with  $\Sigma_\lambda$  by following Eq 1. Setting  $\lambda = 0$ , we can disable the re-weighting, and perform whitening with the biased covariance matrix, while increasing  $\lambda$  gradually incorporates unbiased statistics.

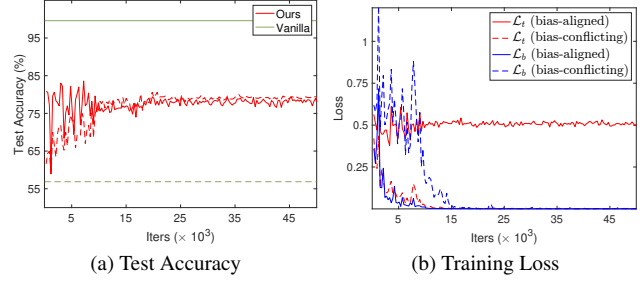


Figure 2. The solid lines and dashed lines in (a) are the test accuracy of the *bias-aligned* and *bias-conflicting* samples, respectively. For comparison, we also report the test accuracy of the *Vanilla* network, which shares the target encoder  $h_t$ . (b) is the illustration of the training loss of  $\mathcal{L}_t$  and  $\mathcal{L}_b$  over the *bias-aligned* and *bias-conflicting* samples.

## 5. Experimental Results

To evaluate the efficacy of the proposed method, we compare the performance with existing methods on one constructed dataset (Corrupted CIFAR10 [34, 40]) and three real-world datasets (Biased FFHQ [30], WaterBirds [49], and Celeb-A [39]). We conduct ablation studies to analyze 1) the contribution of each component in our method and 2) performance variations under different whitening transforms. Due to space constraints, 1) performance comparisons on the WaterBirds dataset, 2) implementation details, and 3) dataset descriptions are provided in Appendix.

**Evaluation metrics.** Following previous studies, we report three types of accuracy: *unbiased*, *bias-conflicting*, and *worst group* [31, 52]. Detailed explanations and equations to compute the metrics are provided in Appendix. To sum up, we can evaluate the utility of algorithms with the *unbiased test* accuracy, and the fairness of algorithms with the *bias-conflicting* and *worst-group* test accuracy.

### 5.1. Controllable Feature Whitening

To evaluate the efficacy of removing linear correlation between the target and bias features which are passed to the last linear layer, we conduct experiments on the bFFHQ dataset. In Figure 2, we illustrate the test accuracy and the training loss of  $\mathcal{L}_t$  and  $\mathcal{L}_b$ . As shown in Figure 2a, the *Vanilla* network, which is highly biased toward  $B$ , performs well only on the *bias-aligned* samples (i.e., young women and old men). In contrast, the performance gap between two groups of our method is significantly reduced, and it indicates that the prediction of our method  $\hat{Y} = g_{wt}(z_{wt})$  is not affected by the bias attribute. Despite the fact that we reuse the biased target encoder  $h_t(\cdot)$  from the *Vanilla* network without update, the proposed whitening module successfully removes the dependency between  $z_{wt}$  and  $B$  as  $z_{wb}$  is trained to predict  $B$ .

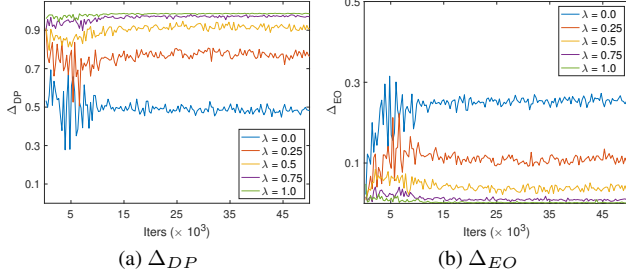


Figure 3. Illustration of  $\Delta_{DP}$  and  $\Delta_{EO}$  with respect to training iterations. It shows that whitening with the unbiased covariance ( $\lambda=1$ ) successfully regularizes  $\Delta_{EO}$ . By contrast, whitening with the biased covariance ( $\lambda=0$ ) successfully regularizes the  $\Delta_{DP}$ .

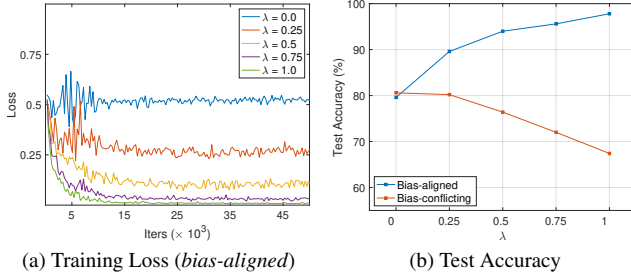


Figure 4. Illustration of training loss of *bias-aligned* samples with respect to training iterations, and test accuracy with respect to  $\lambda$ . As  $\lambda$  is getting larger, training loss of  $\mathcal{L}_t$  on the *bias-aligned* samples is converging better. However, the network tends to be over-fitting due to the sample diversity imbalance between two groups.

Furthermore, as shown in Figure 2b, the training loss of  $\mathcal{L}_t$  for the *bias-aligned* samples does not converge, even though the *bias-aligned* samples are majority in the dataset. By contrast, the training loss of  $\mathcal{L}_t$  over the *bias-conflicting* samples is well-converged, further indicating that the whitening module successfully removes the bias correlated information from  $z_{wt}$  including the task-relevant information. To improve stability of training, we adopt the re-weighting strategy to  $\mathcal{L}_t$  by under-weighting the loss of the *bias-aligned* samples to mimic the loss of unbiased dataset. Empirically, it helps stabilizing, and ablation studies will be provided in Section 5.3.

To preserve task-relevant information, we introduce the Controllable Feature Whitening (CFW), and demonstrate the efficacy by comparing the performance as varying the weight coefficient  $\lambda$ . In Figure 3, we present  $\Delta_{DP}$  and  $\Delta_{EO}$  over the training dataset, computed by Eq 3 and 5. The results allow us to empirically verify whether elimination of linear correlations through CFW can effectively regulate  $\Delta_{DP}$  and  $\Delta_{EO}$ . As we expected, with  $\lambda$  of 0, we observe low  $\Delta_{DP}$  and high  $\Delta_{EO}$ . It indicates that whitening with the biased covariance matrix focuses on removing dependency between  $\hat{Y}$  and  $B$  without conditioning on  $Y$ . By contrast, with  $\lambda$  of 1, we observe high  $\Delta_{DP}$  and low

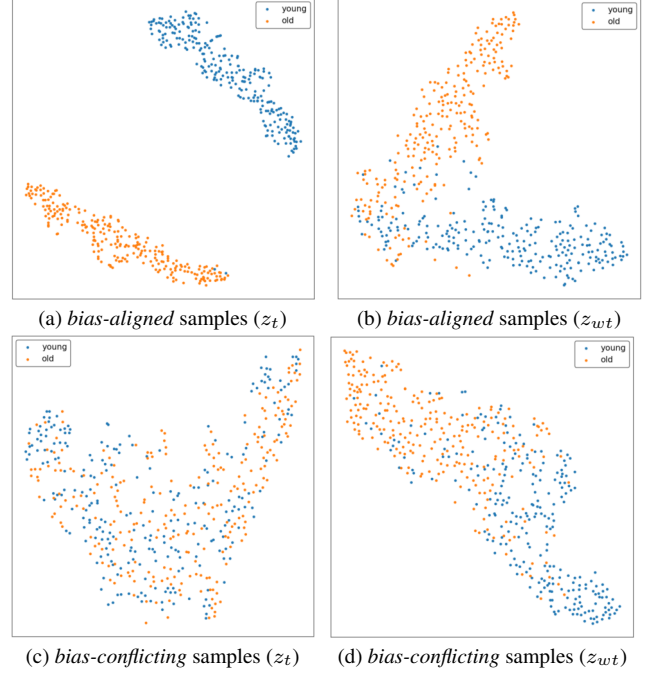


Figure 5. Illustration of the 2D projected target features  $z_t$  and whitened target features  $z_{wt}$  which are extracted with the biased FFHQ dataset. As we expected,  $z_{wt}$  is consistently clustered according to the target attribute (old & young) regardless of the groups, while  $z_t$  is randomly mixed on *bias-conflicting* samples.

$\Delta_{EO}$ . It indicates that whitening with the unbiased covariance matrix removes the dependency between  $\hat{Y}$  and  $B$  conditioning on  $Y$ . These results confirm that removing linear correlations between target and bias features with CFW allows us to regulate smoothly interpolated objective between *demographic parity* and *equalized odds* by adjusting  $\lambda$ .

To further analyze the effect of CFW, we visualize the training loss of  $\mathcal{L}_t$  on the *bias-aligned* samples and test accuracy evaluated on *bias-aligned* samples and *bias-conflicting* samples in Figure 4. As shown in Figure 4a, the training loss of  $\mathcal{L}_t$  in the *bias-aligned* samples successfully converges as  $\lambda$  increases. It indicates that the target information is well-preserved by re-weighting the covariance matrix. However, in Figure 4b, we empirically verify that the performance gap between the *bias-conflicting* and *bias-aligned* samples is growing with increasing  $\lambda$ . Although CFW demonstrates improved accuracy than the *Vanilla* network regardless of  $\lambda$ , it seems that CFW with large  $\lambda$  is prone to over-fitting due to the lack of the sample diversity of the *bias-conflicting* samples. To avoid both target information loss and over-fitting, we empirically set  $\lambda$  to 0.25 based on the results on the bFFHQ dataset. Notably,  $\lambda$  of 0.25 is used in all experiments in Section 5, consistently yielding strong performance across datasets. This demonstrates that our method can be practically considered hyperparameter-free.

| Method                   | Bias Label | Corrupted CIFAR-10 |                   |                   |                   | bFFHQ             |
|--------------------------|------------|--------------------|-------------------|-------------------|-------------------|-------------------|
|                          |            | 0.5%               | 1.0%              | 2.0%              | 5.0%              | 0.5%              |
| Vanilla                  | ✗          | 23.26±0.29         | 26.10±0.72        | 31.04±0.44        | 41.98±0.12        | 56.20±0.35        |
| HEX [53]                 | ✗          | 13.87±0.06         | 14.81±0.42        | 15.20±0.54        | 16.04±0.63        | 52.83±0.90        |
| Rebias [2]               | ✗          | 22.27±0.41         | 25.72±0.20        | 31.66±0.43        | 43.43±0.41        | 59.46±0.64        |
| LfF [40]                 | ✗          | 28.57±1.30         | 33.07±0.77        | 39.91±0.30        | 50.27±1.56        | 62.2±1.0          |
| DisEnt [36]              | ✗          | 29.95±0.71         | 36.49±1.79        | 41.78±2.29        | 51.13±1.28        | 63.87±0.31        |
| SelecMix+L (w/o GT) [22] | ✗          | 39.44±0.22         | 43.68±0.51        | 49.70±0.54        | 57.03±0.48        | 70.80±2.95        |
| EnD [52]                 | ✓          | 22.54±0.65         | 26.20±0.39        | 32.99±0.33        | 44.90±0.37        | 56.53±0.61        |
| LISA [54]                | ✓          | 32.71±1.09         | 38.18±0.90        | 44.15±0.39        | 51.57±0.45        | 64.20±0.53        |
| SelecMix+L (w GT) [22]   | ✓          | 37.02±1.05         | 41.66±1.10        | 48.35±0.99        | 53.47±0.53        | 75.00±0.53        |
| Ours+V                   | ✓          | 32.08±0.32         | 36.13±0.34        | 43.51±0.16        | 53.08±0.22        | 79.8±0.33         |
| Ours+S                   | ✓          | <b>42.51±0.17</b>  | <b>46.87±0.12</b> | <b>50.99±0.16</b> | <b>59.05±0.23</b> | <b>82.77±0.35</b> |

Table 1. Comparison of the unbiased test accuracy (%) on Corrupted CIFAR10 and bias-conflicting test accuracy (%) on bFFHQ. We compare the performance as varying the ratio of *bias-conflicting* samples in the training dataset. We adopted target encoder networks pretrained with Vanilla and SelecMix for ‘Ours+V’ and ‘Ours+S’.

| Method                               | Backbone    | Bias Label | Unbiased          | Bias-conflicting  | Worst-G           |
|--------------------------------------|-------------|------------|-------------------|-------------------|-------------------|
| Target attribute: <i>BlondHair</i>   |             |            |                   |                   |                   |
| Vanilla                              | Res18 (SL)  | ✗          | 70.25±0.35        | 52.52±0.19        | 16.48             |
| LfF [40]                             | Res18 (SL)  | ✗          | 85.43             | 83.40             | -                 |
| LWBC [31]                            | Res18 (SL)  | ✗          | 85.1±0.6          | 82.4±1.4          | 76.6±4.6          |
| Co-Ada [58]                          | Res18 (SL)  | ✗          | -                 | -                 | 78.37             |
| CM [3]                               | Res18 (SL)  | ✗          | -                 | -                 | 81.61             |
| GroupDro [49]                        | Res18 (SL)  | ✓          | 84.24             | 81.24             | -                 |
| CSAD [61]                            | Res18 (SL)  | ✓          | 89.36             | 87.53             | -                 |
| EnD [52]                             | Res18 (SL)  | ✓          | <b>91.21</b>      | 87.45             | -                 |
| Ours                                 | Res18 (SL)  | ✓          | 88.40±0.40        | <b>88.07±0.16</b> | <b>84.03±0.19</b> |
| Vanilla                              | Res18 (SSL) | ✗          | 80.48±0.91        | 66.79±2.20        | 38.5±4.1          |
| LWBC [31]                            | Res18 (SSL) | ✗          | <b>88.90±1.55</b> | <b>87.22±1.14</b> | <b>85.5±1.4</b>   |
| Target attribute: <i>HeavyMakeup</i> |             |            |                   |                   |                   |
| Vanilla                              | Res50 (SL)  | ✗          | -                 | -                 | 47.2              |
| JTT [37]                             | Res50 (SL)  | ✗          | -                 | -                 | 81.1              |
| CNC [59]                             | Res50 (SL)  | △          | -                 | -                 | 88.8±0.9          |
| DFR [33]                             | Res50 (SL)  | △          | 91.3±0.3          | -                 | 88.3±1.1          |
| FMD [8]                              | Res50 (SL)  | △          | 89.73             | -                 | 87.15             |
| SSA [41]                             | Res50 (SL)  | △          | -                 | -                 | 89.8±1.28         |
| GroupDro [49]                        | Res50 (SL)  | ✓          | -                 | -                 | 87.2              |
| LISA [54]                            | Res50 (SL)  | ✓          | -                 | -                 | 89.30             |
| Ours                                 | Res50 (SL)  | ✓          | <b>92.33±0.13</b> | <b>93.24±0.16</b> | <b>91.02±0.17</b> |

Table 2. Comparison of the *unbiased*, *bias-conflicting*, and *worst-group* test accuracy (%) on the Celeb-A. ‘SL’ and ‘SSL’ denote the adoption of supervised and self-supervised models as the backbone network.

For qualitative evaluations, we visualize the 2D projections of the target features  $z_t$  and whitened target features  $z_{wt}$  using t-SNE in Figure 5. These features are extracted from the bFFHQ dataset, and for better clarity, we separately visualize the projections for *bias-aligned* and *bias-conflicting* samples. A highly biased network only performs great on *bias-aligned* samples, but a fair network consistently performs regardless of the groups. As shown in Figure 5b and 5d, the whitened target features  $z_{wt}$  are separable according to the target attribute (old & young) regardless of the groups. On the other hand, as shown in Figure 5a and

5c, the target features  $z_t$  are only accurately separable on the *bias-aligned* samples. It indicates that the proposed method successfully reduces the discrepancy between the groups, and prevents the network predictions from relying on the bias shortcuts.

## 5.2. Classification Results

To compare the performance, we report the *unbiased* and *bias-conflicting* test accuracy on corrupted CIFAR-10 and bFFHQ, respectively, in Table 1. The proposed method demonstrates superior performance, and we note that its effectiveness further enhanced by adopting target encoder that pretrained with SelecMix. We denote our method employing target encoder networks pretrained with *Vanilla* and *SelecMix* as ‘Ours+V’ and ‘Ours+S’, respectively.

In Table 2, we compare the *unbiased*, *bias-conflicting*, and *worst-group* accuracy of the proposed method with other existing methods on Celeb-A. For the *bias-conflicting* and *worst-group* accuracy, our method consistently outperforms other algorithms. For the *unbiased* accuracy, our method shows comparable performance with the best result. In addition, the performance gap between the *unbiased* and *bias-conflicting* accuracy is significantly reduced with our method, indicating that our method improves the fairness without loss of the utility of algorithms. Additionally, in Table 3, we compare the *unbiased* test accuracy and  $\Delta_{EO}$  of the proposed method with other baselines using bias label. The results confirm that our approach achieves both superior fairness and overall performance. Especially, we verify that our method shows significantly better performance than adversarial learning based methods (e.g., GRL and LNL).

## 5.3. Ablation Study

To investigate the contribution of each component of our method, we conduct the ablation study by training ResNet-18 on the bFFHQ under different configurations. LW and

| Method             | T=a / S=m     |             | T=a / S=y     |             | T=b / S=m     |             | T=b / S=y     |             | T=e / S=m     |             | T=e / S=y     |             |
|--------------------|---------------|-------------|---------------|-------------|---------------|-------------|---------------|-------------|---------------|-------------|---------------|-------------|
|                    | $\Delta_{EO}$ | Acc         | $\Delta_{EO}$ | Acc         | $\Delta_{EO}$ | Acc         | $\Delta_{EO}$ | Acc         | $\Delta_{EO}$ | Acc         | $\Delta_{EO}$ | Acc         |
| <i>Vanilla</i>     | 27.8          | 79.6        | 16.8          | 79.8        | 17.6          | <u>84.0</u> | 14.7          | <u>84.5</u> | 15.0          | 83.9        | 12.7          | 83.8        |
| <i>GRL</i> [44]    | 24.9          | 77.2        | 14.7          | 74.6        | 14.0          | 82.5        | 10.0          | 83.3        | 6.7           | 81.9        | 5.9           | 82.3        |
| <i>LNL</i> [29]    | 21.8          | 79.9        | 13.7          | 74.3        | 10.7          | 82.3        | 6.8           | 82.3        | 5.0           | 81.6        | 3.3           | 80.3        |
| <i>FD-VAE</i> [42] | 15.1          | 76.9        | 14.8          | 77.5        | 11.2          | 81.6        | 6.7           | 81.7        | 5.7           | 82.6        | 6.2           | 84.0        |
| <i>MFD</i> [24]    | 7.4           | 78.0        | 14.9          | <u>80.0</u> | 7.3           | 78.0        | 5.4           | 78.0        | 8.7           | 79.0        | 5.2           | 78.0        |
| <i>SupCon</i> [27] | 30.5          | <b>80.5</b> | 21.7          | <b>80.1</b> | 20.7          | <b>84.6</b> | 16.9          | 84.4        | 20.8          | 84.3        | 10.8          | 84.0        |
| <i>FSCl+</i> [43]  | <u>6.5</u>    | 79.1        | <u>12.4</u>   | 79.1        | <b>4.7</b>    | 82.9        | <u>4.8</u>    | 84.1        | <b>3.0</b>    | 83.4        | <b>1.6</b>    | 83.5        |
| <i>Ours</i>        | <b>4.2</b>    | <u>80.3</u> | <b>10.4</b>   | 79.4        | <u>5.0</u>    | 83.5        | <b>4.5</b>    | <b>84.7</b> | <u>3.4</u>    | <b>85.3</b> | <u>2.1</u>    | <b>85.2</b> |

Table 3. Comparison of the top-1 unbiased test accuracy (%) and equalized odds in various scenarios using Celeb-A. Here  $a$ ,  $b$ ,  $e$ ,  $m$ , and  $y$  respectively denote *attractiveness*, *bignose*, *bag-under-eyes*, *male*, and *young*. On the other hand, T and S represent target and sensitive attributes, respectively.

| Method                       | Unbiased (%) | Bias-conflicting (%) | diff. (%)   |
|------------------------------|--------------|----------------------|-------------|
| Vanilla                      | 77.96        | 56.87                | 21.09       |
| + LW                         | 78.5         | 56.2                 | 22.03       |
| + CFW, $\lambda = 0.0$       | 80.0         | <b>80.6</b>          | <b>-0.6</b> |
| + CFW, $\lambda = 0.25$      | <u>83.9</u>  | 78.1                 | 5.8         |
| + CFW, $\lambda = 0.25$ + LW | <b>84.6</b>  | <u>79.8</u>          | <u>4.8</u>  |

Table 4. Ablation study on each component of proposed method. We train and evaluate on the bFFHQ as varying the components. Best performing results are marked in bold, while second-best results are denoted with underlines.

| Method         | Unbiased (%) | Bias-conflicting (%) | diff. (%)   |
|----------------|--------------|----------------------|-------------|
| Vanilla        | 77.96        | 56.87                | 21.09       |
| ZCA            | 83.83        | 72.66                | 11.17       |
| CD             | 74.06        | 71.40                | <b>2.66</b> |
| cNS, T=3       | <u>84.50</u> | 76.80                | 7.7         |
| cNS, T=7       | 83.80        | <u>78.67</u>         | 5.13        |
| Ours: cNS, T=5 | <b>84.6</b>  | <b>79.8</b>          | <u>4.8</u>  |

Table 5. Ablation study on the method to solve the matrix inverse square root. We train and evaluate on the bFFHQ as varying the method. Best performing results are marked in bold, while the second-best results are denoted with underlines.

CFW are abbreviation of loss weighting and Controllable Feature Whitening, respectively. As shown in Table 4, we can verify that all components contribute to achieve the fairness while preserving the utility. Although, we verify that CFW without re-weighting demonstrates the smallest performance gap between two groups, we can optimize the *unbiased* accuracy with the negligible *bias-conflicting* accuracy drop by controlling the weight coefficient  $\lambda$ . Notably, LW does not improve the performance of the *Vanilla* network, while it improves the performance of the proposed method. As we mentioned in Section 4.2, LW helps to stabilize the training by under-weighting the noisy gradients of the *bias-aligned* samples. However, with *Vanilla* network, training loss of both *bias-aligned* and *bias-conflicting* samples converge close to zero even without LW.

We further conduct the ablation study on the whitening module, training with ResNet-18 on the bFFHQ as varying the methods to solve the matrix inverse square root. We adopt three representative methods: ZCA-Whitening [4, 26], Cholesky Decomposition [12], and coupled Newton-Schultz iterations [17, 18], which are denoted as ZCA, CD, and cNI, respectively. For cNI, there is the hyperparameter, iteration number T, which determines the number of iterations to approximate the inverse square root of the matrix. We use the cNI with T of 5, which is suggested by [20, 55]. As shown in Table 5, the fairness and utility are generally improved regardless of which whitening modules are used. However, stochasticity of each modules are different. According to the previous works [21], cNI demonstrates the most stable behavior, and it also outperforms other methods in our experiments. Moreover, we empirically verify that performance improvement is saturated with T of 5.

## 6. Conclusion

In this paper, we propose a novel framework, Controllable Feature Whitening (CFW), to mitigate over-reliance on spurious correlations by removing linear correlations between target features and bias features. Specifically, linear independence ensures that two features cannot be linearly predicted from each other. To enforce this, we whiten the features fed into the last linear classifier, effectively preventing model predictions from relying on bias attributes without requiring intractable modeling of higher-order dependencies. Additionally, we extend our method to achieve two fairness criteria, *demographic parity* and *equalized odds*, by re-weighting the covariance matrix. Although our approach assumes access to bias labels, it demonstrates consistently superior performance across datasets without requiring additional hyperparameter tuning. We validate its effectiveness by achieving state-of-the-art performance on four benchmark datasets: Corrupted CIFAR-10, biased FFHQ, WaterBirds, and Celeb-A.

## References

- [1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 1
- [2] Hyojin Bahng, Sanghyuk Chun, Sangdoon Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 1, 7
- [3] Abhipsa Basu, Saswat Subhajyoti Mallick, et al. Mitigating biases in blackbox feature extractors for image classification tasks. *Advances in Neural Information Processing Systems*, 37:106411–106439, 2025. 7
- [4] Anthony J Bell and Terrence J Sejnowski. The “independent components” of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997. 3, 8
- [5] Dario A Bini, Nicholas J Higham, and Beatrice Meini. Algorithms for the matrix  $p$ th root. *Numerical Algorithms*, 39(4):349–378, 2005. 3
- [6] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019. 3
- [7] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009. 3
- [8] Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. *Advances in Neural Information Processing Systems*, 36:14516–14539, 2023. 7
- [9] Ming-Chang Chiu, Pin-Yu Chen, and Xuezhe Ma. Better may not be fairer: A study on subgroup discrepancy in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4956–4966, 2023. 3
- [10] Yooshin Cho, Hanbyel Cho, Youngsoo Kim, and Junmo Kim. Improving generalization of batch whitening by convolutional unit optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5321–5329, 2021. 3
- [11] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019. 1
- [12] Dariusz Dereniowski and Marek Kubale. Cholesky factorization of matrices in parallel and ranking of graphs. In *International conference on parallel processing and applied mathematics*, pages 985–992. Springer, 2003. 3, 8
- [13] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR, 2021. 3
- [14] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017. 4
- [15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 2, 4
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1
- [17] Nicholas J Higham. Newton’s method for the matrix square root. *Mathematics of computation*, 46(174):537–549, 1986. 3, 8
- [18] Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008. 3, 8
- [19] Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 791–800, 2018. 1
- [20] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4883, 2019. 1, 8
- [21] Lei Huang, Lei Zhao, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. An investigation into the stochasticity of batch whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2020. 8
- [22] Inwoo Hwang, Sangjun Lee, Yunhyeok Kwak, Seong Joon Oh, Damien Teney, Jin-Hwa Kim, and Byoung-Tak Zhang. Selectmix: Debaised learning by mixing up contradicting pairs. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. 7
- [23] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 3
- [24] Sangwon Jung, Donggyu Lee, Taeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12115–12124, 2021. 8
- [25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2
- [26] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018. 3, 8
- [27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 8
- [28] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard

- Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017. 4
- [29] Byungju Kim, Hyunwoo Kim, Kyungso Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020, 2019. 1, 3, 8
- [30] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14992–15001, 2021. 2, 3, 5
- [31] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. *Advances in Neural Information Processing Systems*, 35:18403–18415, 2022. 3, 5, 7
- [32] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 101–110, 2019. 1
- [33] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 4, 7
- [34] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 2, 5
- [35] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017. 3
- [36] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021. 1, 3, 7
- [37] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 3, 7
- [38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2
- [39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5
- [40] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 1, 2, 3, 5, 7
- [41] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022. 7
- [42] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2403–2411, 2021. 8
- [43] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022. 8
- [44] Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198. IEEE, 2018. 8
- [45] Ruggero Ragonese, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. Learning unbiased representations via mutual information backpropagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2729–2738, 2021. 1
- [46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1
- [47] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020. 3
- [48] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Buló, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9471–9480, 2019. 3
- [49] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 2, 5, 7
- [50] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. *arXiv preprint arXiv:2201.03121*, 2022. 2
- [51] Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening and coloring batch transform for gans. *arXiv preprint arXiv:1806.00420*, 2018. 3
- [52] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13508–13517, 2021. 4, 5, 7
- [53] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019. 7
- [54] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution

- robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022. 7
- [55] Chengxi Ye, Matthew Evanusa, Hua He, Anton Mitrokhin, Tom Goldstein, James A Yorke, Cornelia Fermüller, and Yiannis Aloimonos. Network deconvolution. *arXiv preprint arXiv:1905.11926*, 2019. 3, 8
- [56] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 1
- [57] 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, 2018. 4
- [58] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022. 7
- [59] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 3, 7
- [60] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 3
- [61] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15002–15012, 2021. 1, 3, 7
- [62] Zhuotun Zhu, Lingxi Xie, and Alan L Yuille. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596*, 2016. 1