This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Self-supervised debiasing using low rank regularization

Geon Yeong Park[1], Chanyong Jung[1], Sangmin Lee[2], Jong Chul Ye[1,2,3*] Sang Wan Lee[1,4*]
[1]Bio and Brain Engineering, [2]Mathematical Sciences,
[3]Kim Jaechul Graduate School of AI, [4]Brain and Cognitive Sciences
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{pky3436, jcy132, leeleesang, jong.ye, sangwan}@kaist.ac.kr

Abstract

Spurious correlations can cause strong biases in deep neural networks, impairing generalization ability. While most existing debiasing methods require full supervision on either spurious attributes or target labels, training a debiased model from a limited amount of both annotations is still an open question. To address this issue, we investigate an interesting phenomenon using the spectral analysis of latent representations: spuriously correlated attributes make neural networks inductively biased towards encoding lower effective rank representations. We also show that a rank regularization can amplify this bias in a way that encourages highly correlated features. Leveraging these findings, we propose a self-supervised debiasing framework potentially compatible with unlabeled samples. Specifically, we first pretrain a biased encoder in a self-supervised manner with the rank regularization, serving as a semantic bottleneck to enforce the encoder to learn the spuriously correlated attributes. This biased encoder is then used to discover and upweight bias-conflicting samples in a downstream task, serving as a boosting to effectively debias the main model. Remarkably, the proposed debiasing framework significantly improves the generalization performance of self-supervised learning baselines and, in some cases, even outperforms state-of-the-art supervised debiasing approaches.

1. Introduction

While modern deep learning solves several challenging tasks successfully, a series of recent works [16, 18, 20] have reported that the high accuracy of deep networks on in-distribution samples does not always guarantee low test error on out-of-distribution (OOD) samples, especially in the context of spurious correlations. Existing studies [3, 38, 50] suggest that the deep networks can be potentially biased to the spuriously correlated attributes, or dataset bias, which are misleading statistical heuristics that are closely correlated but not causally related to the target label.

These catastrophic pitfalls of dataset bias have facilitated the development of debiasing methods, which can be roughly categorized into approaches: (**1**) leveraging annotations of spurious attributes, i.e., bias label [27, 46, 49, 56]; (**2**) presuming specific type of bias, e.g., color and texture [5, 17, 53]; or (**3**) without using explicit kinds of supervisions on dataset bias [30, 31, 35, 39, 60].

While substantial advances have been made in this regard, these approaches still fail to address the problem: how to train a debiased classifier by fully exploiting unlabeled samples lacking *both* bias and target label. More specifically, while the large-scale unlabeled dataset can be potentially biased towards spuriously correlated sensitive attributes, e.g., ethnicity, gender, or age [1, 2], current existing debiasing frameworks are not designed to deal with this real-world unsupervised settings. Here we also confirm that most supervised debiasing frameworks suffer from performance degradation in the low-labeled data setting. Moreover, recent works have suggested that self-supervised learning might not be sufficient to deal with OOD generalization [10, 19, 44] when dataset bias remains after data augmentation.

To tackle this issue, we first make a series of empirical observations that allow us to examine the fundamental difference between biased and unbiased representations. Interestingly, we found that spurious correlations suppress the effective rank [45] of latent representations, which severely deteriorates the semantic diversity of representations and leads to the degradation of feature discriminability. Another notable aspect of our findings is that the intentional increase of feature redundancy amplifies "prejudice" in neural networks. To be specific, as we enforce the correlation among latent features to regularize the effective rank of representations (i.e., rank regularization), the accuracy on bias-conflicting samples quickly declines while the model still performs reasonably well on the bias-aligned [1] samples.

[1]The *bias-aligned* samples refer to data with a strong correlation between (potentially latent) spurious features and target labels. The *bias-conflicting* samples refer to the opposite cases where spurious correlations do not exist.

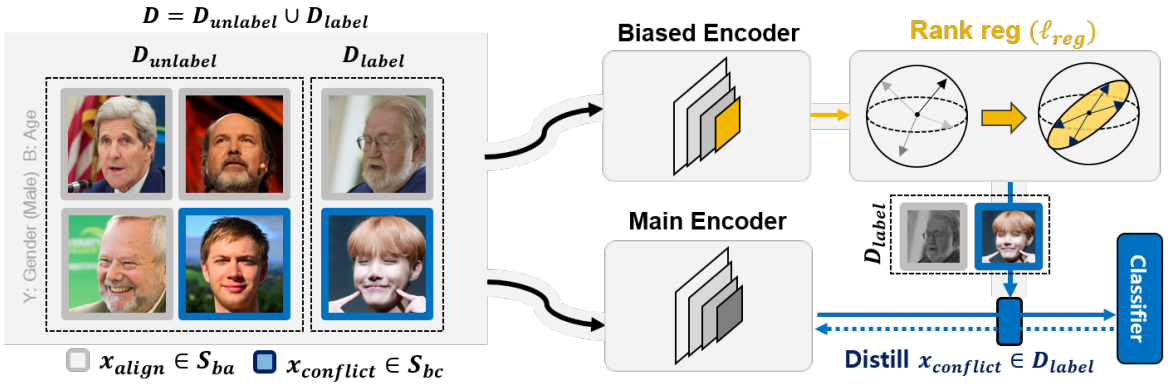


Figure 1. **Concept.** Based on the bias-rank relationship (Section 2), we introduce a novel debiasing framework centered on rank regularization, which intentionally amplifies spurious correlation by enforcing feature components to be *entangled* with both spurious and invariant attributes.

Based on these observations, we propose a novel debiasing framework that can utilize both labeled *and* unlabeled biased samples with rank regularization. The proposed method is fully compatible with both supervised and self-supervised scenarios, where such compatibility arises from the rank regularization that does *not* rely on any labels. Specifically, for a supervised (self-supervised) setting, we train 1) a biased classifier (encoder) with rank regularization, which serves as a semantic bottleneck limiting the semantic diversity of feature components, and 2) the main classifier (encoder) with standard (self-)supervised learning approaches. The biased model affords us the leverage to uncover spurious correlations and identify bias-conflicting samples in a downstream task.

Our work is the first to unveil the bias-rank relationships and introduce an effective debiasing strategy to exploit potentially unlabeled data samples. We demonstrate the effectiveness of the proposed debiasing framework with various challenging real-world biased datasets, including MultiCMNIST [34], biased Chest X-ray databases, UTKFace, CelebA, etc., in both a supervised and self-supervised scenario. These experiments show that our method significantly outperforms other self-supervised baselines, and even state-of-the-art supervised debiasing methods in some cases.

2. Low-rank bias of biased representations

2.1. Preliminaries

Throughout the paper, we denote $x \in \mathbb{R}^m$ and $y \in \mathcal{Y}$ as m-dimensional input sample and its corresponding predicting label, respectively. Then we denote $X = \{x_k\}_{k=1}^n$ as a batch of n samples from a dataset which is fed to an encoder $f_\theta : \mathbb{R}^m \to \mathbb{R}^d$, parameterized by θ. Then we construct a matrix $Z \in \mathbb{R}^{n \times d}$ where each ith row is the output representations of the encoder $f_\theta(x_i)^T$ for $x_i \in X$. For every analysis in this section, we use Z as our latent representations, where the neural backbone of the encoder may vary as simple convolutional networks, ResNet-18, or ViT [15] (Experimental details provided in Appendix C.1 and D).

To evaluate the semantic diversity of given representation matrix, we introduce *effective rank* [45] which is a widely used metric to measure the effective dimensionality of matrix and analyze the spectral properties of features in neural networks [4, 6, 23, 42]:

Definition 1. *Given the matrix $X \in \mathbb{R}^{m \times n}$ and its singular values $\{\sigma_i\}_{i=1}^{\min(m,n)}$, the effective rank ρ of X is defined as the shannon entropy of normalized singular values:*

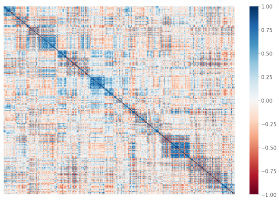
$$\rho(X) = -\sum_{i=1}^{\min(m,n)} \bar{\sigma}_i \log \bar{\sigma}_i, \tag{1}$$

where $\bar{\sigma}_i = \sigma_i / \sum_k \sigma_k$ is i-th normalized singular value. Without loss of generality, we omit the exponentiation of $\rho(X)$ as done in [45].

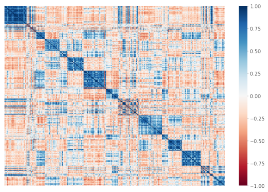
Effective rank is also referred to as spectral entropy where its value is maximized when the singular values are all equal and minimized when a top singular value dominates relative to all others. Recent works [12, 13] have revealed that the discriminability of representations resides on wide range of eigenvectors since the rich discriminative information for the classification task cannot be transmitted by only few eigenvectors with top singular values. Thus from a spectral analysis perspective, effective rank quantifies how diverse the semantic information encoded by each eigenfeature is, which is closely related to the feature discriminability across target label categories. In the rest of paper, we interchangeably use effective rank and rank by following prior works.

2.2. Spectral analysis of the bias-rank relationships

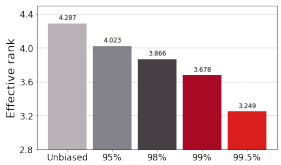
We now present experiments showing that the deep networks may tend to encode lower-rank representations in the pres-



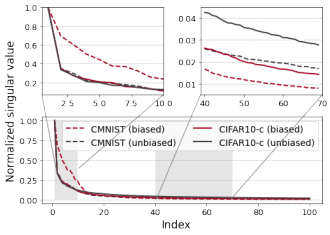
(a) Unbiased corr.



(b) Biased corr.



(c) Effective rank



(d) SVD analysis

Figure 2. Empirical analysis on rank reduction phenomenon. For every analysis, we used the output Z of the encoder (Sec. 2.1). (**a**, **b**): Hierarchically clustered auto-correlation matrix of unbiased and biased representations (Bias ratio=99%). (**c**): Effective rank with color bias. 'Unbiased' represents the case training model with perfectly unbiased dataset, i.e. random color for each sample. (**d**): SVD analysis with max-normalized singular values. Top 100 values are shown in the figure (total: 256).

ence of stronger spurious correlations. To arbitrarily control the degree of spurious correlations, we introduce synthetic biased datasets, Color-MNIST (CMNIST) and Corrupted CIFAR-10 (CIFAR-10C, [21]), with color and corruption bias types, respectively. We define the degree of spurious correlations as the ratio of bias-aligned samples included in the training set, or bias ratio, where most of the samples are bias-aligned in the context of strong spurious correlations.

Figure 2c shows that the rank of latent representations from a penultimate layer of the simple convolutional classifier decreases as the bias ratio increases in CMNIST. We provide similar rank reduction results of CIFAR-10C with ResNet-18 and ViT in the Appendix C.1. We further compare the correlation matrix of biased and unbiased latent representations in the penultimate layer of biased and unbiased classifiers, respectively. In Figure 2a and 2b, we observe that the block structure in the correlation matrix is more evident in the biased representations after the hierarchical clustering, indicating that the features become highly correlated which may limit the semantic diversity of networks. To investigate the rank reduction phenomenon in-depth, we compare the normalized singular values of biased and unbiased representations. We conduct singular value decomposition (SVD) on the feature matrices of both biased and unbiased classifiers and plot the singular values normalized by the spectral norm of the corresponding matrix. Figure 2d shows that the top few normalized singular values of biased representations are similar to or even greater than those of unbiased representations. However, the remaining majority of singular values decay significantly faster in biased representations, greatly weakening the informative signals of eigenvectors with smaller singular values and deteriorating feature discriminability [12, 13].

2.3. Rank regularization

Motivated from the aforementioned rank reduction phenomenon, we ask an opposite-directional question: "Can we intentionally amplify the prejudice of deep networks by *maximizing* the redundancy between the components of latent representations?". If the feature components are extremely correlated, the corresponding representations may exhibit most of its spectral energy along the direction of one singular vector. For this case, effective rank may converge to 0. In other words, our goal is to design a *semantic bottleneck* of representations that restricts the semantic diversity of feature vectors. To implement the bottleneck in practice, motivated from Figure 2b, we compute the auto-correlation matrix of the output of encoder.

Let $\bar{Z}$ denote the mean-centered representations Z along the batch dimension. The normalized auto-correlation matrix $C \in \mathbb{R}^{d\times d}$ of $\bar{Z}$ is defined as follow:

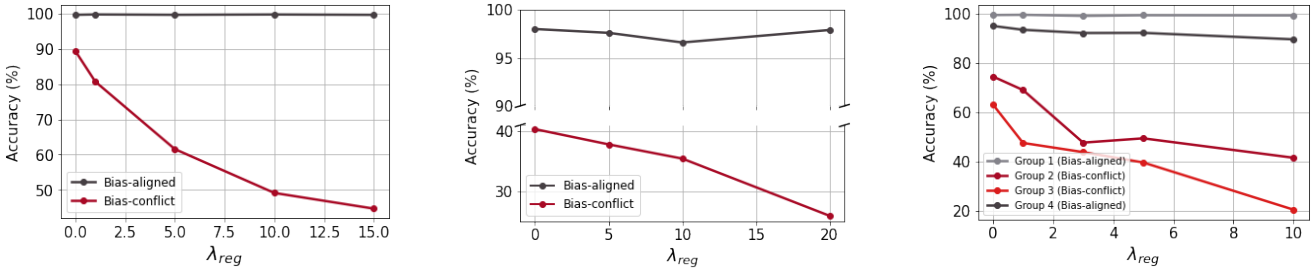
$$C_{i,j} = \frac{\sum_{b=1}^{n} \bar{Z}_{b,i}\bar{Z}_{b,j}}{\sqrt{\sum_{b=1}^{n} \bar{Z}_{b,i}^2}\sqrt{\sum_{b=1}^{n} \bar{Z}_{b,j}^2}} \quad 1 \le \forall i,j \le d, \quad (2)$$

where b is an index of sample and i, j are index of each vector dimension. Then we define our regularization term as the negative of a sum of squared off-diagonal terms in C:

$$\ell_{reg}(X;\theta) = -\sum_{i}\sum_{j\neq i} C_{i,j}^2, \quad (3)$$

where we refer to it as the *rank loss*. Note that the target labels on X is *not* used at all.

Analysis of rank-regularized networks. To investigate the impacts of rank regularization in deep neural networks, we construct the classification model by combining the linear classifier $f_W : \mathbb{R}^d \to \mathbb{R}^c$ parameterized by $W \in \mathcal{W}$ on top of the encoder f_θ, where $c = |\mathcal{Y}|$ is the number of classes. Then we trained models by cross entropy loss ℓ_{CE} combined with $\lambda_{reg}\ell_{reg}$, where $\lambda_{reg} > 0$ is a Lagrangian multiplier. We use CMNIST, CIFAR-10C, and Waterbirds dataset [51], and evaluate the trained models on an unbiased test set following [30, 39]. After training models with varying the hyperparameter λ_{reg}, we compare bias-aligned and bias-conflict accuracy, which are the average accuracy on bias-aligned and bias-conflicting samples in the unbiased test set, respectively, for CMNIST and CIFAR-10C. Test accuracy on every individual data group is reported for Waterbirds. Figure 3 shows that models suffer more from poor OOD generalization as trained with larger λ_{reg}. The average



(a) CMNIST (b) CIFAR-10C (c) Waterbirds

Figure 3. (**a**, **b**): Bias-conflict and Bias-aligned accuracy on CMNIST and CIFAR-10C (Bias ratio=95%). (**c**): Group accuracy on Waterbirds. Detailed simulation settings are in the Appendix D.

	CMNIST		CIFAR-10C	
	P (%)	R (%)	P (%)	R (%)
ERM	85.59	19.76	52.03	0.06
+ Rank reg	**98.83**	**95.91**	**71.39**	**51.43**

(a) CMNIST, CIFAR-10C

Metrics	ERM	JTT	Rank reg
Precision (%)	37.84	48.95	**54.77**
Recall (%)	11.67	48.75	**55.01**

(b) Waterbirds

Table 1. Precision (P) and Recall (R) of bias-conflicting samples. (**a**): Bias-conflicting samples are identified in the error set of ERM model trained with and without rank regularization (Bias ratio=95% for both datasets). (**b**): Bias-conflicting samples are similarly identified by ERM, JTT, and the proposed biased model in Waterbirds dataset.

accuracy on bias-conflicting groups is significantly degraded, while the accuracy on bias-aligned groups is maintained to some extent. It implies that rank regularization may force deep networks to focus on spurious attributes.

Minority mining performance. Table 1a and 1b support that the biased models with strong regularization can effectively probe out the bias-conflicting samples in the training set. Specifically, we train a biased classifier with rank regularization and distill an error set E of misclassified training samples as bias-conflicting samples proxies. As reported in Table 1a, we observe that our biased classifier is relatively robust to the unintended memorization of bias-conflicting samples [47] in contrast to the standard models trained by Empirical Risk Minimization (ERM). Moreover, Table 1b shows that the proposed rank regularization improves the precision and recall of identified bias-conflicting samples compared to JTT [35]. Detailed simulation settings are in the Appendix D.

Reconstruction of biased representations. To understand the relationship between rank regularization and spurious correlations more deeply, we visualize the pretrained representations with varying degrees of bias. We first trained deep networks on: (**a**) unbiased CMNIST (random background color), (**b**) biased CMNIST (bias ratio=95%) without rank regularization and (**c**) with rank regularization ($\lambda_{reg} = 50$). Then, we train the auxiliary decoder, which reconstructs the bias-conflicting images from the freezed latent representations of each pretrained network. Results show that rank regularization may cause the representation to lose information on complex invariant features, resulting in a loss of feature discriminability and informative signals. While both digit and color are well reconstructed with biased representations (**b**), the decoder fails to reconstruct bias-conflicting images from the (**c**) biased representations pretrained with rank regularization. The foreground digit is blurred, and its class changes following the color-digit assignment in Figure 4d.

These observations afford us some key insights into rank regularization: First, the rank-regularized representation may lose its information on complex invariant features (i.e., shape and style of the foreground digit), specifically undermining the feature discriminability and informative signals. Second, the limited semantic diversity makes it harder to identify the true underlying independent generative factors for multidimensional data; instead, it may encode feature components *entangled* with both spurious and invariant attributes as the digit class of the reconstructed image is erroneously determined by the background color in 4c.

Multiple bias attributes. To further investigate the generalizability of rank regularization, we evaluate the biased representations with Multi-Color MNIST (MultiCMNIST) dataset [34], which is similar to the CMNIST but have two bias attributes: left and right background colors. We set bias ratio=99% for the left color and bias ratio=95% for the right color, i.e., the left color is a more salient bias than the right color (Dataset details are provided in Appendix D).

Table 2 shows that the rank regularization successfully biases the model w.r.t both bias attributes, while LfF [39]

(a) Unbiased (b) Biased (c) Biased with rank reg. (d) Examples.

Figure 4. Randomly selected reconstructed images from representations with varying degrees of bias. First and third row correspond to the input bias-conflicting images. Second and fourth row correspond to the reconstructed images. Reconstructed from (**a**) unbiased representations, (**b**) biased representations, and (**c**) biased representations with rank regularization (bias ratio=95% in **b, c**). (**d**) Examples of bias-aligned CMNIST images.

Table 2. (**a**) Test accuracy (%) on MultiCMNIST. Lower is better for this results. BC for bias-conflicting, and BA for bias-aligned. Bias ratio=99(%) for left color, and 95(%) for right color. $\lambda_{reg} = 50$ is used for rank regularization. (**b**) Debiasing results. Higher is better for this results. Baseline results are from [34]. $\lambda_{up} = 50$ is used for upweighting in the proposed framework (λ_{up}: a manual rescaling weight given to each identified bias-conflicting samples in cross entropy loss). Pseudo-code and experimental details are provided in Appendix A and D, respectively.

Idx	Left color	Right color	(**a**) Biased accuracy (%)			(**b**) Debiased accuracy (%)		
			ERM	LfF [39]	**Rank reg.**	LfF	DebiAN [34]	**Ours**
(1)	BA	BC	100.0	100.0	100.0	99.6	100.0	100.0
(2)	BA	BC	96.6	98.8	41.6	4.7	95.6	97.0
(3)	BC	BA	29.3	3.2	8.7	98.6	76.5	79.1
(4)	BC	BC	7.6	1.3	6.1	5.1	16.0	18.3
(1) ∼ (4) average acc.			58.38	50.83	**39.1**	52.0	72.0	**73.6**

completely fails to amplify the right color bias, i.e. less salient bias, as shown in the second row (Biased accuracy part). This leads to the abnormal debiasing results of LfF as shown in Table 2 where it records unbalanced accuracy for the left- and right-color-bias-conflicting samples. In contrast, the proposed framework shows superior performance by simply upweighting the misclassified bias-conflicting proxies, as done in [35].

Taken together, these results indicate that the rank regularization encourages the network to focus more on spurious correlations in a way that minimizes semantic diversity and *entangles* invariant and spurious features [41], which is a fundamentally different mechanism compared to the LfF [39] with its easy-to-learn assumption. More details on the upweighting strategy will be provided in Section 3 and pseudo-code in Appendix A.

3. DeFund: Debiasing framework with unlabeled data

Motivated by the observations in Section 2, we propose a self-supervised debiasing framework with unlabeled data, coined DeFund (Debiasing Framework with Unlabeled Data). A notable distinction from previous studies [5, 60] lies in the proposed framework's ability to effectively harness unlabeled data for learning biased representations. This is achieved through the application of self-supervised learning and rank regularization techniques.

The proposed framework is composed of two stages: We first train the biased encoder, which can be potentially adopted to detect the bias-conflicting samples in a downstream task, along with the main encoder by self-supervised learning, both without any labels. After pretraining, we identify the bias-conflicting samples in the downstream task using linear evaluation protocol [9, 40]. This set of samples serves as a boosting to debias the main model.

Notation. We denote $f_\theta^{bias} : \mathcal{X} \to \mathbb{R}^d$ and $f_\phi^{main} : \mathcal{X} \to \mathbb{R}^d$ as biased encoder and main encoder parameterzied by $\theta \in \Theta$ and $\phi \in \Theta$, respectively, where d is the dimensionality of latent representations. Then we can compute the rank loss in (3) with introduced encoders and given batch $\{x_k\}_{k=1}^N$ with size N. Let $f_{W_b}^{cls} : \mathbb{R}^d \to \mathbb{R}^C$ be a single-layer classifier parameterized by $W_b \in \mathcal{W}$ which is placed on top of biased encoder f_θ^{bias}, where $C = |\mathcal{Y}|$ is the number of classes. We similarly define the linear classifier $f_{W_m}^{cls}$ for the main encoder. Then we refer to $f^{bias} : \mathcal{X} \to \mathbb{R}^C$ as biased model, where $f^{bias}(x) = f_{W_b}^{cls}\big(f_\theta^{bias}(x)\big), \forall x \in \mathcal{X}$. We similarly define the main model f^{main} as $f^{main}(x) = f_{W_m}^{cls}\big(f_\phi^{main}(x)\big), \forall x \in \mathcal{X}$. While the projection networks [9] are employed as well, we omit the notations because they are not engaged in classification.

Stage 1. Training a biased encoder. To train the biased encoder f_θ^{bias}, we revisit the proposed rank regularization term (3) in context of instance discrimination task. Building upon the observations in Section 2.3, we conjecture that rank regularization may amplify bias in self-supervised learning as well by entangling invariant and spurious features. Based on these intuitions, we apply rank regularization directly to the output of the base encoder, which encourages each feature component to be highly correlated. From these applications, several noteworthy observations have emerged: (**a**) The representation becomes more biased as it is trained with stronger regularization (Appendix C.1). (**b**) While the overall performance may be upper-bounded due to the constraint on effective dimensionality [25], the bias-conflict accuracy

is primarily sacrificed compared to the bias-aligned accuracy (Section 4).

Stage 2. Debiasing downstream tasks. After training the biased encoder, our goal is to debias the main model, which was pretrained using standard self-supervised learning methods on the same dataset. Here, assume that we have an ideal pretrained main encoder of which each output component corresponds to the latent factor of data variation [63]. While this ideal encoder should seamlessly adapt to downstream classification tasks, if most downstream task samples are bias-aligned, they may misguide the model to upweight spuriously correlated latent factors, leading to a biased solution despite well-generalized representations. We refer to this problem as the biased downstream application (Related analysis in Appendix C.1).

The above contradiction elucidates the importance of bias-conflicting samples, which serve as counterexamples of spuriously correlated feature components, thereby preventing the alleged involvement of such components in prediction. Based on these intuitions, we introduce a novel debiasing protocol that probes and upweights bias-conflicting samples to find and fully exploit feature components independent of spurious correlations. We apply our framework on two scenarios: linear evaluation and semi-supervised learning.

Linear evaluation. To validate our hypothesis on the biased downstream application, we conduct linear evaluation [40, 61] following the conventional protocol of self-supervised learning. Specifically, a linear classifier is trained on top of unsupervised pretrained representations by using target labels of training samples. After training a linear classifier $f^{cls}_{W_b}$ with pretrained biased encoder f^{bias}_{θ} given the whole training set $D = \{(x_k, y_k)\}_{k=1}^{N}$ with size N, an error set E of misclassified samples and corresponding labels is regarded as bias-conflicting pairs. Then we train a linear classifier $f^{cls}_{W_m}$ on intentionally freezed representations of main encoder f^{main}_{ϕ} by upweighting the identified samples in E with $\lambda_{up} > 0$. The loss function for *debiased* linear evaluation is defined as follows:

$$\ell_{debias}(D; W_m) = \lambda_{up} \sum_{(x,y)\in E} \ell(x, y; W_m) + \sum_{(x,y)\in D\setminus E} \ell(x, y; W_m),$$

where we use cross entropy loss for $\ell : \mathcal{X}\times\mathcal{Y}\times\mathcal{W} \to \mathbb{R}^{+}$. Note that the target labels are only used in training linear classifiers after pretraining.

Note that the debiased linear evaluation is not meant to compete directly with other supervised baselines. Instead, it aims to: (**a**) examine the potential origin of the failure in OOD generalization, (**b**) provide a rough estimate of the potential improvement achievable with frozen latent representations, and (**c**) compare with standard self-supervised baselines and identify the optimal learning algorithms, e.g. SimCLR [9], for training the main encoder.

Semi-supervised learning. We further compare our method directly to other supervised debiasing methods in the context of semi-supervised learning. Here we assume that the training dataset includes only a small amount of labeled data combined with a large amount of unlabeled data. As in linear evaluation, we train a linear classifier on top of the biased encoder by using labeled samples. After obtaining an error set E of misclassified samples, we finetuned the whole main model by upweighting the identified samples in E with λ_{up}. Note that supervised baselines are restricted to using only a small fraction of labeled samples, while the proposed approach benefits from the abundant unlabeled samples during pre-training of the biased encoder (Pseudo-code in the Appendix section A).

4. Results

4.1. Methods

Dataset. We evaluate several supervised and self-supervised baselines on **MIMIC-CXR + NIH** [32], **UTKFace** [62] and **CelebA** [36] in which prior work reported poor generalization performance due to spurious correlations (Dataset details in Appendix).

For MIMIC-CXR + NIH, we mixed the MIMIC-CXR [26] and NIH [55] following [32] where the target categories are `no finding` and `pneumonia`. Most `pneumonia` images are collected from MIMIC-CXR, while most `no finding` images are from NIH. In other words, the biases come from systematic differences in data sources, where the classifier may erroneously rely on spurious radiographic features tied to variations in data acquisition pipelines [14] instead of true pathological indicators (Example images in Figure 8).

For UTKFace, we conduct binary classifications using (`Gender`, `Age`) and (`Race`, `Gender`) as (target, spurious) attribute pair, which we refer to UTKFace (age) and UTKFace (gender), respectively. For CelebA, we consider (`HeavyMakeup`, `Male`) and (`Blonde Hair`, `Male`) as (target, spurious) attribute pairs, which are referred to CelebA (makeup) and CelebA (blonde), respectively. The results of CelebA (blonde) are reported in Appendix C.4. Following [22, 39], we report bias-conflict accuracy together with unbiased accuracy, which is evaluated on the explicitly constructed validation set. We exclude the dataset in Figure 3 based on the observations that the SimCLR models are already invariant w.r.t spurious attributes.

Baselines. We mainly target baselines consisting of recent advanced self-supervised learning methods, SimCLR [9], VICReg [7], and SimSiam [11], which can be categorized into contrastive (SimCLR) and non-contrastive (VICReg, SimSiam) methods. We further report the performance of vanilla networks trained by ERM, and other supervised debiasing methods such as LNL [27], EnD [49], and

Table 3. (Supervised learning) Bias-conflict and unbiased accuracy (%) on MIMIC-CXR + NIH. Each ✓marker represents whether the model requires information on dataset bias. Bias ratio=10%.

Accuracy	LNL ✓	EnD ✓	LfF ✗	JTT ✗	CVaR DRO ✗	ERM ✗	SimCLR ✗	**Ours** ✗
Conflict	$43.8_{\pm 0.5}$	$50.4_{\pm 2.3}$	$25.2_{\pm 2.1}$	$47.9_{\pm 0.2}$	$44.6_{\pm 0.5}$	$41.7_{\pm 1.2}$	$35.5_{\pm 1.3}$	$\mathbf{56.8}_{\pm 1.7}$
Unbiased	$68.1_{\pm 1.0}$	$\mathbf{71.8}_{\pm 1.4}$	$60.8_{\pm 0.2}$	$68.9_{\pm 1.0}$	$65.8_{\pm 1.2}$	$67.8_{\pm 1.0}$	$62.0_{\pm 1.4}$	$69.8_{\pm 0.2}$

Table 4. (Linear evaluation) Bias-conflict and unbiased test accuracy (%) evaluated on UTKFace and CelebA. Models requiring information on target class or dataset bias in the (pre)training stage are denoted with ✓in columns Y and B, respectively. Our results are marked in bold to highlight the improvements compared to the mainly interested self-supervised learning baselines (Gray rows).

Model	Y	B	UTKFace (age)		UTKFace (gender)		CelebA (makeup)	
			Conflict	Unbiased	Conflict	Unbiased	Conflict	Unbiased
LNL	✓	✓	$45.8_{\pm 0.6}$	$72.6_{\pm 0.3}$	$73.1_{\pm 1.6}$	$84.9_{\pm 0.8}$	$55.9_{\pm 2.1}$	$76.0_{\pm 0.6}$
EnD	✓	✓	$45.3_{\pm 0.9}$	$72.2_{\pm 0.2}$	$75.5_{\pm 1.1}$	$85.5_{\pm 0.4}$	$57.3_{\pm 2.4}$	$76.4_{\pm 1.4}$
JTT	✓	✗	$63.8_{\pm 0.9}$	$69.4_{\pm 1.3}$	$71.2_{\pm 0.3}$	$77.6_{\pm 0.4}$	$62.4_{\pm 1.2}$	$74.7_{\pm 0.8}$
CVaR DRO	✓	✗	$45.7_{\pm 2.0}$	$71.4_{\pm 0.3}$	$68.6_{\pm 1.0}$	$81.0_{\pm 0.8}$	$58.0_{\pm 1.7}$	$76.5_{\pm 0.6}$
ERM	✓	✗	$45.4_{\pm 2.1}$	$71.0_{\pm 1.2}$	$65.7_{\pm 1.4}$	$79.5_{\pm 0.6}$	$54.2_{\pm 0.2}$	$74.1_{\pm 1.4}$
SimSiam	✗	✗	$28.2_{\pm 0.9}$	$62.6_{\pm 0.7}$	$48.5_{\pm 1.0}$	$69.8_{\pm 0.7}$	$39.9_{\pm 0.6}$	$66.7_{\pm 0.6}$
VICReg	✗	✗	$32.3_{\pm 0.6}$	$64.6_{\pm 0.3}$	$51.0_{\pm 1.4}$	$71.3_{\pm 0.7}$	$48.6_{\pm 0.6}$	$71.9_{\pm 0.2}$
SimCLR	✗	✗	$36.4_{\pm 1.5}$	$66.3_{\pm 0.6}$	$56.3_{\pm 0.2}$	$74.2_{\pm 0.2}$	$46.9_{\pm 1.0}$	$69.8_{\pm 0.4}$
DeFund	✗	✗	$\mathbf{59.5}_{\pm 0.8}$	$\mathbf{70.6}_{\pm 0.8}$	$\mathbf{63.7}_{\pm 2.0}$	$\mathbf{74.9}_{\pm 0.9}$	$\mathbf{58.4}_{\pm 0.6}$	$\mathbf{73.1}_{\pm 1.0}$

upweighting-based algorithms, JTT [35] and CVaR DRO [31], which can be categorized into methods that leverage annotations on dataset bias (LNL, EnD) or not (JTT, CVaR DRO).

Optimization setting. Both bias and main encoder is pretrained with SimCLR [9] for 100 epochs on UTKFace, and 20 epochs on CelebA, respectively, using ResNet-18, Adam optimizer and cosine annealing learning rate scheduling [37]. We use a MLP with one hidden layer for projection networks as in SimCLR. All the other baseline results are reproduced by tuning the hyperparameters and optimization settings using the same backbone architecture. We report the results of the model with the highest bias-conflicting test accuracy over those with improved unbiased test accuracy compared to the corresponding baseline algorithms, i.e., SimCLR for ours (More experimental details in Appendix D).

4.2. Evaluation results

Supervised learning. To quantify the effectiveness of the rank regularization in-depth, we first consider a standard supervised debiasing scenario as similarly done in Table 2. For a MIMIC-CXR + NIH dataset, we found that the proposed framework outperforms other supervised baselines with respect to bias-conflict accuracy. Table 14 in the Appendix shows that the rank-regularized networks effectively discover the bias-conflicting samples which are consistent with Table 1a, 1b, and 2.

Linear evaluation. We also found that DeFund outperforms every self-supervised baseline by a large margin in a linear evaluation protocol, including SimCLR, SimSiam and VICReg, with respect to both bias-conflict and unbiased accuracy (Table 4). Moreover, in some cases, DeFund even outperforms ERM models or supervised debiasing approaches regarding bias-conflict accuracy. Note that there is an inherent gap between ERM models and self-supervised baselines, roughly 8.7% on average. Moreover, we found that non-contrastive learning methods generally perform worse than the contrastive learning method. This warns us against training the main model using a non-contrastive learning approach, while it may be a viable option for the biased model. Results of the proposed framework with non-contrastive learning methods are provided in the Appendix section C.5.

Semi-supervised learning. To compare supervised and self-supervised methods in a more practical and fair scenario, we randomly sample 10% of the labeled CelebA training dataset. The remaining 90% samples are treated as unlabeled ones and engaged only in pretraining encoders for self-supervised baselines. Labeled samples are provided equally to both supervised and self-supervised methods.

Table 5. (Semi-supervised learning) Accuracy results (%) on CelebA. Label fraction= 10%.

Accuracy	CelebA (Makeup)							CelebA (Blonde)	
	LNL	EnD	JTT	CVaR DRO	ERM	SimCLR	**DeFund**	JTT	**DeFund**
Conflict	$55.7_{\pm 1.4}$	$55.3_{\pm 1.5}$	$51.5_{\pm 1.9}$	$55.6_{\pm 1.5}$	$51.5_{\pm 1.1}$	$50.5_{\pm 4.7}$	$\mathbf{60.5}_{\pm 0.4}$	$70.6_{\pm 1.0}$	$\mathbf{75.1}_{\pm 0.8}$
Unbiased	$75.6_{\pm 0.5}$	$\mathbf{76.2}_{\pm 0.8}$	$71.4_{\pm 1.3}$	$75.7_{\pm 1.0}$	$73.1_{\pm 0.3}$	$71.6_{\pm 1.9}$	$75.6_{\pm 0.2}$	$78.8_{\pm 1.7}$	$\mathbf{85.8}_{\pm 0.3}$

Remarkably, Table 5 and Table 16 in Appendix show that the proposed framework outperforms other state-of-the-art supervised debiasing methods. Existing upweighting protocols, such as JTT, fail to prevent deep networks from memorizing minority counterexamples. However, the proposed framework can fully utilize unlabeled samples with contrastive learning to prevent memorization. Existing bias-conflicting sample mining algorithms may be affected by the implicit bias of overparameterized networks, but this is unlikely to happen with the proposed framework since it only trains a simple linear classifier on top of a frozen biased encoder to identify such samples.

Method	UTKFace (age)		UTKFace (gender)		CelebA (makeup)	
	Conflict	Unbiased	Conflict	Unbiased	Conflict	Unbiased
SimCLR	36.4	66.3	56.3	74.2	46.9	69.8
+ Rank reg	26.6	61.3	50.9	70.3	43.9	68.3
+ Upweight	53.0	64.6	58.3	74.5	50.1	70.4
DeFund	**59.5**	**70.6**	**63.7**	**74.9**	**58.4**	**73.1**

(a) Ablation study

Method	UTKFace (age)		UTKFace (gender)		CelebA (makeup)	
	Precision	Recall	Precision	Recall	Precision	Recall
SimCLR	68.31	44.63	**33.36**	39.59	52.25	28.23
DeFund	**68.67**	**75.94**	29.98	**50.93**	**55.29**	**32.46**

(b) Precision and recall

Table 6. (**a**) Ablation study on introduced modules. (**b**) Precision and recall (%) of bias-conflicting samples identified by SimCLR and our biased model. Both case used linear evaluation.

Ablation study. To quantify the extent of performance improvement achieved by each introduced module, we compared the linear evaluation results of (**a**) vanilla SimCLR, (**b**) SimCLR with rank regularization, (**c**) SimCLR with upweighting error set E of the main model, and (**d**) DeFund. Note that (**c**) does not use a biased model at all. Table 6a shows that every module plays an important role in OOD generalization. Considering that the main model is already biased to some extent, we found that bias-conflict accuracy can be improved even without a biased model, where the error set E of the biased model further boosts the generalization performance. We also measures the precision and recall of identified bias-conflicting samples in E, finding that the biased model detects more diverse bias-conflicting samples than the baseline (Table 6b). The improvement of recall in CelebA may seem marginal, but it is significant given the larger number of samples compared to UTKFace.

Computational costs. Our framework is computationally affordable as it only trains the linaer classifier (linear eval.) or finetune networks with a few epochs, e.g., about 30 epochs for UTKFace in debiasing stage. Self-supervised pre-training and linear evaluation takes 19.3 and 4.5 minutes with a NVIDIA GeForce RTX 2080Ti, respectively.

5. Conclusion

We present a novel solution to the challenging self-supervised debiasing, an important problem that has received little attention so far. We (**a**) unveil the inductive bias towards encoding low effective rank representations in the presence of spurious correlations. Based on this, we (**b**) design a rank regularization that amplifies the feature redundancy by reducing the spectral entropy of latent representations. Then we (**c**) design a debiasing framework empowered by the biased model pretrained with abundant unlabeled samples.

Acknowledgments

This research was supported by National Research foundation of Korea(NRF) (**RS-2023-00262527**), Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711137899, KMDF_PR_20200901_0015), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT, Ministry of Science and ICT) (No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation), (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)), (No. RS-2023-00233251, System3 reinforcement learning with high-level brain functions), Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government (24ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System), [N01230878, Development of Beyond X-verse Core Technology for Hyper-realistic interactions by Synchronizing the Real World and Virtual Space], a grant of the KAIST-KT joint research project through AI2XL Laboratory, Institute of convergence Technology, funded by KT (Project No.3, Project title: Genie Brain).

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.

[2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Hyojin Bahng, Sanghyuk Chun, Sangdoo 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.

[6] Aristide Baratin, Thomas George, César Laurent, R Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. Implicit regularization via neural feature alignment. In *International Conference on Artificial Intelligence and Statistics*, pages 2269–2277. PMLR, 2021.

[7] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[8] Tiffany Tianhui Cai, Jonathan Frankle, David J Schwab, and Ari S Morcos. Are all negatives created equal in contrastive instance discrimination? *arXiv preprint arXiv:2010.06682*, 2020.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[12] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019.

[14] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[17] Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. *Advances in Neural Information Processing Systems*, 34:27356–27368, 2021.

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[19] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. *arXiv preprint arXiv:2010.08377*, 2020.

[20] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

[21] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[22] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

[23] Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *arXiv preprint arXiv:2103.10427*, 2021.

[24] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[25] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[26] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[27] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[28] 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.

[29] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip

Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021.

[30] 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.

[31] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[32] Jiaxuan Li, Duc Minh Vo, and Hideki Nakayama. Partition-and-debias: Agnostic biases mitigation via a mixture of biases-specific experts. *arXiv preprint arXiv:2308.10005*, 2023.

[33] Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14970–14979, 2021.

[34] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pages 270–288. Springer, 2022.

[35] 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.

[36] 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.

[37] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[38] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

[39] 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.

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[41] Geon Yeong Park, Sangmin Lee, Sang Wan Lee, and Jong Chul Ye. Training debiased subnetworks with contrastive weight pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7929–7938, 2023.

[42] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

[43] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[44] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[45] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[46] 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.

[47] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[48] Afrina Tabassum, Muntasir Wahed, Hoda Eldardiry, and Ismini Lourentzou. Hard negative sampling strategies for contrastive representation learning. *arXiv preprint arXiv:2206.01197*, 2022.

[49] 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.

[50] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[51] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[52] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[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.

[54] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[55] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[56] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[57] Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. Increasing robustness to spurious correlations using forgettable examples. *arXiv preprint arXiv:1911.03861*, 2019.

[58] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[59] Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pages 12356–12367. PMLR, 2021.

[60] 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.

[61] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[62] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

[63] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.