Difficulty-based Sampling for Debiased Contrastive Representation Learning

Taeuk Jang, Xiaoqian Wang*
Purdue University
465 Northwestern Ave, West Lafayette, IN 47907, USA
{jang141@,joywang}@purdue.edu

Abstract

Contrastive learning is a self-supervised representation learning method that achieves milestone performance in various classification tasks. However, due to its unsupervised fashion, it suffers from the false negative sample problem: randomly drawn negative samples that are assumed to have a different label but actually have the same label as the anchor. This deteriorates the performance of contrastive learning as it contradicts the motivation of contrasting semantically similar and dissimilar pairs. This raised the attention and the importance of finding legitimate negative samples, which should be addressed by distinguishing between 1) true vs. false negatives; 2) easy vs. hard negatives. However, previous works were limited to the statistical approach to handle false negative and hard negative samples with hyperparameters tuning. In this paper, we go beyond the statistical approach and explore the connection between hard negative samples and data bias. We introduce a novel debiased contrastive learning method to explore hard negatives by relative difficulty referencing the bias amplifying counterpart. We propose triplet loss for training a biased encoder that focuses more on easy negative samples. We theoretically show that the triplet loss amplifies the bias in self-supervised representation learning. Finally, we empirically show the proposed method improves downstream classification performance.

1. Introduction

The key idea of contrastive learning [4, 5, 31] is to learn the representation that projects samples from the same class to be closer to each other than samples from different classes in the embedding space. To ensure this property in an unsupervised manner, we randomly draw a sample (anchor, $\mathbf{x}^a$) and enforce it to stay closer to its own augmentations (positive samples, $\mathbf{x}^+$) and be apart from the other samples (negative samples, $\mathbf{x}^-$) also randomly drawn from the same training dataset. Such approach achieves superior performance over conventional supervised classification methods in various tasks, such as object detection [12, 39, 43] and natural language processing [28].

Recent works study sampling methods to draw good-quality positive and negative samples to train effective self-supervised contrastive learning models. Various augmentation and positive sampling techniques are developed to boost the performance and generalization. For example, random noise perturbations [4, 7, 39] are adopted in the computer vision domain to preserve semantic information such as random cropping, random noise injection, and tilting. Unlike positive sampling, finding *legitimate* negative samples is not a trivial problem. First, negative samples are not guaranteed to have a different class from the anchor [8, 19] due to the unsupervised fashion of contrastive learning. Thus, the debiasing method [8] was proposed to address this false negatives problem by decomposing the marginal data distribution. Second, finding hard negative samples, *i.e.,* hard to distinguish from the anchor, is crucial as they are more informative [36]. Supervised contrastive learning [19] validated the importance of hard negative mining. However, this has been rarely studied in the literature.

In supervised learning, *e.g.,* classification, some works observed hard samples are related to data bias. Because some models tend to be misled by some correlation between biasing attributes and target labels, such as *texture, color*, and *background* in image classification [2, 25], and *race* and *gender* in face recognition [17], samples against such correlation are likely to be hard samples. For instance, in the animal classification task, if most bird images in the training set are assumed to have *sky* as a background instead of others, *sky* would be strongly correlated with the class *bird*. However, birds may also exist in other backgrounds, such as water, rock, etc. We can consider these birds in the background other than the sky as *bias-conflicting* samples. Then, it is natural to emphasize bias-conflicting samples (birds on water) more than bias-aligned ones (birds in the sky) for better performance and generalization as they are more informative. From the contrastive learning view-

*Corresponding author.

point, these bias-conflicting samples are likely to be hard to distinguish from the anchor (*e.g.,* frog on water) and are naturally linked to hard negatives in the representation space.

To address this, some methods [2, 20] specified the bias based on empirical observations on the task. For example, CNN is known to be biased towards the texture [10]. Bahng *et al.* [2] proposed an adversary that focuses exclusively on texture and limits the size of the receptive field of the convolutional layer to predict the target. However, it is almost infeasible to pre-define the bias attributes for each task, and also, the debiasing would be limited to the specified attributes. Recent studies [25, 26, 30] proposed to emphasize bias-conflicting samples by up-weighting hard samples without pre-defined bias information in the supervised classification task. Yet, this approach is limited to classification tasks. Despite the importance of finding hard negatives [36], little attention is paid to finding bias-conflicting samples in self-supervised representation learning methods.

Unlike the previous studies, we delve into the question: *what makes a sample hard negative or easy negative in self-supervised learning?* To the best of our knowledge, few studies in contrastive learning have been done from this perspective. In this work, we propose a novel contrastive learning method to effectively find hard negative samples from the data bias perspective. We employ triplet loss [38] to learn bias-amplified representations in a self-supervised manner. In Section 5.2, we theoretically show that minimizing triplet loss enforces a model to focus on easy samples and ignores hard samples. Along with the biased model, we train the debiased model based on the relative difficulty of each sample by measuring relative distance between the representation from two models and the anchor as the surrogate of sample difficulty.

The contribution of this work is summarized as follows:

1. We propose a debiased contrastive learning method that addresses two types of biases: hard vs. easy negatives and true vs. false negatives.

2. We introduce triplet loss to amplify the bias in the representation space, which serves as an effective surrogate for learning the relative difficulty of samples in self-supervised contrastive learning.

3. We empirically validate that our learned representation achieves higher accuracy and reduced bias in downstream tasks compared with related methods in image and tabular data classification.

2. Related Work

Contrastive Representation Learning

Contrastive learning is a self-supervised representation method. Recent studies in contrastive learning [14, 43] empirically showed that it improves both performance and robustness in downstream tasks in various domains, including NLP and computer vision. Contrastive learning enforces a sampled data (anchor) to stay closer to its own augmentations (positive samples) and apart from other samples (negative samples) randomly drawn from the dataset. To address the limitation of self-supervised contrastive learning, the debiasing method [8] was introduced to correct the sampling of negative data points. For better negative mining, the importance sampling technique [36] was proposed to further accentuate hard negative samples. Mixing strategy [18] and adversarial attack [44] were proposed to improve the quality of negative samples at the feature level. Furthermore, a recent study [15] suggests canceling out false negatives to minimize their negative impact.

On top of that, supervised method [19] shows that the network can weigh more on hard positive/negative samples and become robust to the perturbations. Even in the supervised setting, Park *et al.* [33] point out that contrastive learning suffers from underfitting problem on bias-conflicting samples when the data distribution is skewed. However, few studies have been done on the connection between negative sampling and data bias in self-supervised contrastive learning. This is an important problem as most real-world scenarios have underlying biases (whether they are known or unknown), and they have a potential risk of discrimination or reliability issues in downstream tasks.

Debiasing in Classification Task

Meanwhile, mitigating bias has been actively studied in the classification task. Given the pre-defined biases, *e.g.,* color or texture, the debiased models are trained adversarially with the bias-targeted models [2, 42]. However, the bias features are not always obvious, which makes the process demanding, and it is hard to generalize to a different setting. Some recent approaches [3, 24, 26, 30] explore the intuition that bias-conflicting samples are likely to be the samples that are hard to learn. Nam *et al.* [30] proposed to learn a debiased model by examining the relative difficulty of a sample compared with the biased models. However, these methods have limited use cases as they use require ground truth label to find bias conflicting samples.

To the best of our knowledge, this work is the first that employs bias-amplifying approach in a self-supervised manner to train the biased encoder. The biased encoder is utilized to debias the representation learning process, which leads to better generalization and bias mitigation.

3. Motivation

Per the goal of contrastive learning, we want positive samples to be close to the anchor and negative samples to be apart from the anchor. In the data bias perspective, the *bias-conflicting* negative sample would be semantically similar

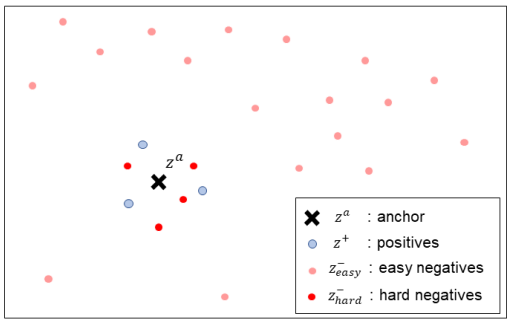


Figure 1. Illustration of the motivating example. For better performance and mitigating bias, we want to emphasize more on hard negative samples $\mathbf{z}^-_{hard}$ (dark red, which are closer to the anchor $\mathbf{z}^a$ (black cross)) than easy negatives $\mathbf{z}^-_{easy}$ (light red).

to the anchor even though it is from a different class. As a result, it is likely to be embedded close to the anchor, *i.e.,* hard negatives, as the cosine similarity is proportional to the distance between two samples as $||\mathbf{x}-\mathbf{y}||_2^2 \propto -2\mathbf{x}^\intercal\mathbf{y}$. Similarly, the *bias-conflicting* positive samples are likely to be far from the anchor, *i.e.,* hard positives, respectively. In contrast, *bias-aligned* negative (*resp.* positive) samples would be embedded far (*resp.* close) to the anchor and easily distinguished. In the following context, we use bias-conflicting/aligned and hard/easy interchangeably.

To address the biased representation learning, we need to focus on hard positive/negative samples. However, positive samples are simple augmentations of the anchors in most cases, we would focus on hard negative samples. Figure 1 illustrates the anchor $\mathbf{z}^a$ and corresponding hard negative samples $\mathbf{z}^-_{hard}$, easy negatives $\mathbf{z}^-_{easy}$, and positive samples $\mathbf{z}^+$ in embedded space.

Suppose there exists a biased encoder E_b that amplifies the bias in learned representation than the *non-bias-amplifying* encoder E. Then the embedding $\mathbf{z}^-_{hard,b}$ of a hard negative sample $\mathbf{x}^-{}_{hard}$ from E_b would be embedded closer to the anchor than the embedding $\mathbf{z}^-_{hard}$ from the other encoder E, *i.e.,* $D(\mathbf{z}^a, \mathbf{z}^-_{hard,b}) < D(\mathbf{z}^a, \mathbf{z}^-_{hard})$, where $D(\cdot,\cdot)$ is a distance between two points. Oppositely, the embedding of an easy negative sample from the biased encoder E_b would be further apart from the anchor than the embedding from the other encoder, *i.e.,* $D(\mathbf{z}^a, \mathbf{z}^-_{easy,b}) > D(\mathbf{z}^a, \mathbf{z}^-_{easy})$. If we have such bias-amplifying encoder, we can estimate the difficulty of negative samples by comparing the relative distance between the anchor and negative sample from the two encoders.

4. Problem Definition

Here, we consider the image classification task as an example. Let $\mathbf{x} \in \mathbb{R}^{w\times h} \sim p(\mathbf{x})$ be an image. We denote $\mathbf{z} = E(\mathbf{x})$ as an embedding of $\mathbf{x}$ by an encoder E.

4.1. Self-Supervised Contrastive Learning [4]

Contrastive learning takes three components as the input: an anchor $\mathbf{x}^a \sim p(\mathbf{x})$, N positive samples $\{\mathbf{x}^{+(i)}\}_{i=1}^N \sim p^+(\mathbf{x}^+|\mathbf{x}^a)$ that are semantically similar to $\mathbf{x}$, and M negative samples $\{\mathbf{x}^{-(j)}\}_{j=1}^M \sim p(\mathbf{x})$. The positive samples are usually augmented from $\mathbf{x}$, while the negative samples are randomly drawn from the same training set as the anchor is drawn from. Contrastive learning aims to learn a representation $E : \mathcal{X} \to \mathcal{Z} \subset \mathcal{S}^{d-1}/t$ that maps $\mathbf{x}$ to a hypersphere $\mathcal{S}^{d-1}/t \subset \mathbb{R}^d$ of radius $1/t$, where t is a scaling hyperparameter and also known as temperature, such that the similar pair $(\mathbf{x}^a, \mathbf{x}^+)$ stay close and all dissimilar pairs $\{(\mathbf{x}^a, \mathbf{x}^-{}_j)\}_{j=1}^M$ are far apart in the normalized embedded space $\mathcal{Z}$ provided by E. To achieve the goal, E is optimized to minimize the following:

$$\mathbb{E}_{\mathbf{x}^a,\mathbf{x}^+,\mathbf{x}^-}\left[-\log\frac{e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^+)}}{e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^+)}+\sum_{j=1}^M e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^{-(j)})}}\right].$$

4.2. Debiased Contrastive Learning

Finding *legitimate* negative samples takes a significant role in training an effective contrastive learning model, which should be addressed by distinguishing between 1) true vs. false negative; 2) easy vs. hard negative.

4.2.1 Debias False Negative [8]

Not all negatives are necessarily true negative samples in the unsupervised setting, *i.e.,* there can be negative samples that have the same class as the anchor. To compensate this, we decompose the marginal data distribution $p(\mathbf{x})$ as

$$p(\mathbf{x}) = \tau^+ p^+(\mathbf{x}) + \tau^- p^-(\mathbf{x}), \tag{1}$$

where $\tau^+ = p(y)$ is class probability which is assumed to be uniform, and $\tau^- = 1-\tau^+$ be the probability of observing different classes from y. Then naive $p^-(\mathbf{x})$ can be substituted by $(p(\mathbf{x}) - \tau^+ p^+(\mathbf{x}))/\tau^-$. Thus, instead of the distance term *w.r.t.* the negative sample in the second term of the denominator of (4.1), we have

$$\mathbb{E}_{\mathbf{x}^a,\mathbf{x}^+,\mathbf{x}^-}\left[-\log\frac{e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^+)}}{e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^+)}+g}\right], \tag{2}$$

where

$$g = \frac{1}{1-\tau^+}\left(\mathbb{E}_{\mathbf{x}^-\sim p(\mathbf{x})}[e^{E(\mathbf{x})^\intercal E(\mathbf{x}^-)}] - \tau^+\mathbb{E}_{\mathbf{x}^+\sim p^+(x)}[e^{E(\mathbf{x})^\intercal E(\mathbf{x}^-)}]\right), \tag{3}$$

to debias wrongly selected negative samples when taking the limit $M \to \infty$.

4.2.2 Hard Negative Mining [36]

It is not a trivial problem to effectively distinguish hard/easy negatives. Importance sampling technique is proposed to focus on hard negative samples. Instead of $p(\mathbf{x})$, they propose to sample from $q(\mathbf{x}) \propto e^{\beta E(\mathbf{x})^\intercal E(\mathbf{x}^-)} \cdot p(\mathbf{x})$, which is an unnormalized distribution proportional to the similarity between two samples. With importance sampling techniques, we can express negative samples as

$$\mathbb{E}_{\mathbf{x}^- \sim q(\mathbf{x})}\left[e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^-)}\right] = \mathbb{E}_{\mathbf{x}^- \sim p(\mathbf{x})}\left[e^{(\beta+1)E(\mathbf{x}^a)^\intercal E(\mathbf{x}^-)}/Z_\beta\right],$$

with the partition function Z_β of $q(\mathbf{x})$ as $Z_\beta = \mathbb{E}_p\left[e^{\beta E(\mathbf{x}^a)^\intercal E(\mathbf{x}^-)}\right]$, and the same applies to $q^+(\mathbf{x})$. By substituting two expectations over $p(\mathbf{x})$ and $p^+(\mathbf{x})$ in (3) with expectations over $q(\mathbf{x})$ and $q^+(\mathbf{x})$, we draw negative samples proportional to the similarity between the anchor.

However, with this importance sampling technique, we also have to sample positive samples with high similarity with the anchor, *i.e.,* easy positives or bias-aligning positives, for debiasing. Moreover, the debiasing term of q^+ in the last term in (3) with q^+ forces easy positive to be farther apart from the anchor as we minimize the contrastive loss. This makes trade-off between hard negatives and easy positives and can potentially risk the training unstable. Also, they have an additional similarity weighting hyperparameter β, which requires extra effort to tune.

5. Difficulty based Debiasing Methods

In this section, we demonstrate the debiasing methods which leverage relative difficulty by comparing with the biased model. We first introduce the existing supervised method and the shortcoming of the works. Then we describe how our proposed method addresses the limitations.

5.1. Bias Amplification in Supervised Setting

Here we consider the binary classification task for simplicity, but this can be generalized to multi-class classification. Denote $\mathbf{x}$ as the input and y as the binary target label. The classifier C outputs probabilistic prediction, *i.e.,* $C(\mathbf{x}) \in [0, 1]$. We train a debiased classifier C_d by comparing a relative loss with a biased classifier C_b.

To train a biased classifier C_b, recent works [25, 30] employ generalized cross entropy (GCE) [45] as:

$$\mathcal{L}_{GCE}(C_b(\mathbf{x};\theta_b), y) = \frac{1 - \big(C_b(\mathbf{x};\theta_b) - (1-y)\big)^q}{q},$$

where $q \in (0, 1]$ as a hypterparameter and θ_b is learnable parameter of C_b. Based on the gradient of GCE, we can easily show that GCE loss weights more on samples where the softmax prediction $C_b(\mathbf{x})$ matches the target label y:

$$\frac{\partial \mathcal{L}_{GCE}(C_b(\mathbf{z};\theta_b), y)}{\partial \theta} = \big(C_b(\mathbf{z};\theta_b) - (1-y)\big)^q \cdot \frac{\partial \mathcal{L}_{CE}}{\partial \theta}.$$

Thus, training C_b to minimize $\mathcal{L}_{GCE}$ makes the classifier to focus more on easy samples.

To train a debiased classifier C_d, the relative difficulty is adopted, which is to compare the prediction $C_b(\mathbf{x})$ with $C_d(\mathbf{x})$. To be specific, we weigh more on the samples with higher CE loss from the output of C_b than that of C_d as C_b is more inattentive to the hard samples due to the nature of GCE loss. In other words, the relative difficulty for each sample $\mathbf{x}$ can be formulated as:

$$w(\mathbf{x}) = \frac{\mathcal{L}_{CE}(C_b(\mathbf{x};\theta_b), y)}{\mathcal{L}_{CE}(C_d(\mathbf{x};\theta_d), y) + \mathcal{L}_{CE}(C_b(\mathbf{x};\theta_b), y)},$$

by comparing CE loss from C_b and C_d. Minimizing weighted cross entropy loss as (4) enforces the classifier C_b to concentrate on hard samples.

$$\underset{\theta_d}{\operatorname{argmin}}\, w(\mathbf{x};\theta_b,\theta_d) \cdot \mathcal{L}_{CE}(C_d(\mathbf{x};\theta_d), y) \tag{4}$$

However, such debiasing method is inapplicable in self-supervised contrastive learning methods as $\mathcal{L}_{GCE}$ requires target label y.

5.2. Debiased Contrastive Learning with Difficulty-based Sampling

In previous works, GCE loss was employed to amplify the bias in classification tasks, however, this is not applicable in self-supervised learning. Inspired by the approach in Section 5.1, we extend the concept of referencing a biased model to self-supervised learning and propose debiased contrastive learning with relative difficulty. Unlike existing contrastive learning methods, our model consists of two encoders E_d and E_b. We want E_d to effectively handle hard negative samples that can mitigate bias and E_b to intentionally amplify bias that focuses only on easy samples.

Inspired by previous supervised learning works in identity classification [38] and supervised contrastive learning [19], we employ triplet loss to amplify the bias in E_b. However, previously triplet loss is used in a supervised manner, which does not apply in our setting without labels. To address this, we introduce the self-supervised triplet loss as:

$$\mathcal{L}_{tri} = \mathbb{E}\big[||E_b(\mathbf{x}^a) - E_b(\mathbf{x}^+)||_2^2 - ||E_b(\mathbf{x}^a) - E_b(\mathbf{x}^-)||_2^2\big]. \tag{5}$$

When we update E_b to minimize $\mathcal{L}_{tri}$ with gradient descent,

$$\nabla_{\theta_b}\mathcal{L}_{tri} = \mathbb{E}\Big[2\Delta^{+\intercal}\nabla\big(E_b(\mathbf{x}^a) - E_b(\mathbf{x}^+)\big) - 2\Delta^{-\intercal}\nabla\big(E_b(\mathbf{x}^a) - E_b(\mathbf{x}^-)\big)\Big], \tag{6}$$

where θ_b is the learnable parameter of E_b and

$$\Delta^+ = E_b(\mathbf{x}^a) - E_b(\mathbf{x}^+), \quad \Delta^- = E_b(\mathbf{x}^a) - E_b(\mathbf{x}^-),$$

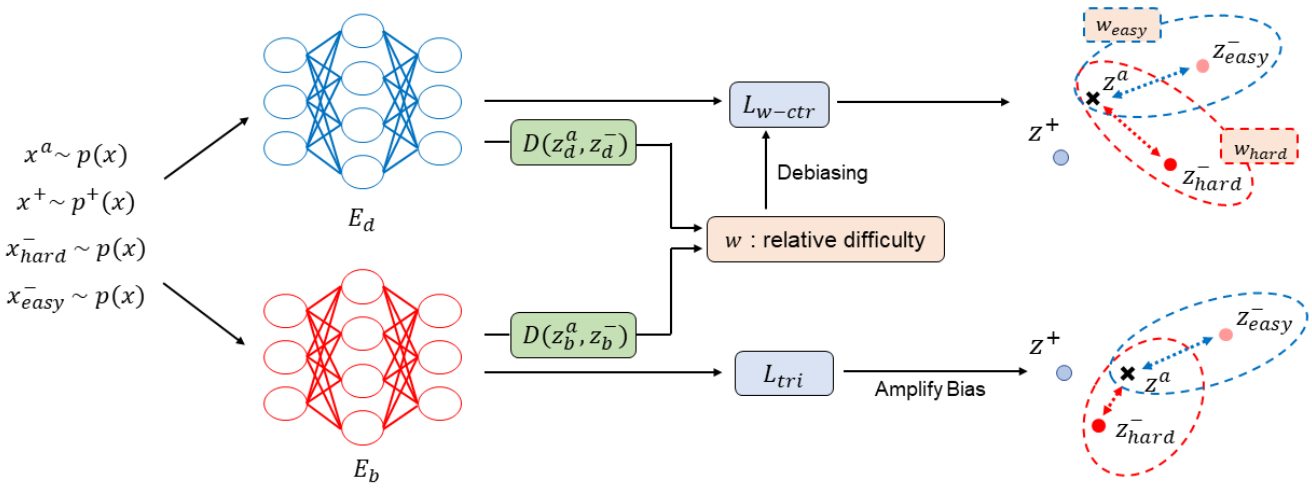


Figure 2. The overall structure of the proposed debiased contrastive learning method. Given the input (x^a, x^+, x^-), identically structured biased encoder E_b and debiased encoder E_d project the data into the latent space. The biased encoder E_b focuses heavily on easy samples that amplifies bias by minimizing $\mathcal{L}_{tri}$. To train the debiased encoder E_d, we leverage the relative difficulty by referencing the biased representation from E_b. We compare the distance between $D(\mathbf{z}^a_d, \mathbf{z}^-_d)$ and $D(\mathbf{z}^a_b, \mathbf{z}^-_b)$ to compute the relative difficulty w. The smaller the relative distance $D(\mathbf{z}^a_b, \mathbf{z}^-_b)$ compared with $D(\mathbf{z}^a_d, \mathbf{z}^-_d)$, the higher weight is assigned to the negative sample, *i.e.,* $w \gg 1$. The debiased encoder E_d is trained to minimize $\mathcal{L}_{w-ctr}$ that up-weights hard negative samples by multiplying this difficulty function w.

measures the distance between the anchor and positive/negative samples.

Suppose we have two negative samples $\mathbf{x}^{-(1)}$ and $\mathbf{x}^{-(2)}$. If $\mathbf{x}^{-(1)}$ is closely located in embedding space, *i.e.,* $\Delta^- \approx 0$, the triplet loss would ignore the sample. In contrast, even if $\mathbf{x}^{-(2)}$ is well distinguished from the anchor, it can dominate the update as it is proportional to the distance Δ^-. Thus, when we train an encoder with triplet loss, we can amplify the bias as easy negatives are more heavily weighted than hard negatives.

Someone might question that Δ^+ would focus on hard positives, which contradicts our intention to amplify the bias. However, it is worth mentioning that $\mathbf{x}^+$ is an augmentation of $\mathbf{x}^a$ with small perturbations *e.g.,* additive random noise, cropping, etc. Thus in practice, Δ^+ are small enough that Δ^- dominates the gradient of the triplet loss in (6). Moreover, we debias the false negative and replace the second term in (5) similar to (2) as we adopt triplet loss in self-supervised setting:

$$\frac{1}{\tau^-}\Big(E_{\mathbf{x}^- \sim p}\big[||E_b(\mathbf{x}^a) - E_b(\mathbf{x}^-)||_2^2\big] - \tau^+ \mathbb{E}_{\mathbf{x}^+ \sim p^+}\big[||E_b(\mathbf{x}^a) - E_b(\mathbf{x}^+)||_2^2\big]\Big).$$

Now, we introduce how to learn debiased encoder E_d parallelly with E_b. Unlike traditional contrastive learning in (4.1), we weight each negative sample differently by relative difficulty as

$$w\big((\mathbf{z}^a_d, \mathbf{z}^-_d), (\mathbf{z}^a_b, \mathbf{z}^-_b)\big) = 1 + \frac{\tilde{D}(\mathbf{z}^a, \mathbf{z}^-_d)}{\tilde{D}(\mathbf{z}^a, \mathbf{z}^-_d) + \tilde{D}(\mathbf{z}^a, \mathbf{z}^-_b)},$$

where $\tilde{D}(\cdot, \cdot)$ is euclidean distance normalized in a batch. In other words, the normalized euclidean distance between representation from E_i can be expressed as

$$\tilde{D}(\mathbf{z}^a_i, \mathbf{z}^-_i) = \frac{D(\mathbf{z}^a_i, \mathbf{z}^-_i)}{\max_{(\mathbf{x}^a, \mathbf{x}^-) \in \mathcal{B}} D(E_i(\mathbf{x}^a), E_i(\mathbf{x}^-)},$$

where $\mathbf{z}^a_i = E_i(\mathbf{x}^a), \forall(\mathbf{x}^a, \mathbf{x}^-) \in \mathcal{B}$. We normalize the distance to limit the maximum distance of a pair from each encoder E_i to 1. This allows us to compare the relative distance of a pair between two encoders E_b and E_d in different latent spaces. Moreover, measuring relative difficulty by simply comparing the outcome from different encoders allows us to avoid demanding hyperparameter tuning.

This relative difficulty function $w \approx 2$ when negative sample is hard negative, *i.e.,* $\mathbf{x}^-{}_d \gg \mathbf{x}^-{}_b$, and $w \approx 1$ when the negative sample is easy negative, *i.e.,* $\mathbf{x}^-{}_d \ll \mathbf{x}^-{}_b$. In other words, we can penalize more on bias-conflicting samples by multiplying relative difficulty function w as:

$$\mathbb{E}\left[-\log \frac{e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^+)}}{e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^+)} + w(\mathbf{z}^a, \mathbf{z}^-_b, \mathbf{z}^-_d) e^{E(\mathbf{x}^a)^\intercal E(\mathbf{x}^-)}}\right].$$

Here we use the anchor as a reference point to measure relative difficulty instead of directly comparing the distance between two embeddings of negative samples. This is because the latent space is the hypersphere with a radius $1/t$

Algorithm 1 Debiased Contrastive Learning via Difficulty-based Sampling

Input training set $\mathcal{X}$, learning rate α, epoch n
Output Debiased encoder E_d and biased encoder E_b.
Randomly Initialize parameters θ_d and θ_b for encoders E_d and E_b, respectively.
while not converge **do**
  for t = 1, 2, $\ldots, n$ **do**
    for $\mathbf{x}_{t_i}$ in the t-th mini-batch $\mathcal{X}_t$ **do**
      1. Draw positive/negative samples for each anchor $\mathbf{x}^a_{t_i}$ and embed to latent space

$$\mathbf{x}^+{}_{t_i} \in Aug(\mathbf{x}^a_{t_i}), \quad \mathbf{x}^-{}_{t_i} \in \mathcal{X}_t \setminus \{\mathbf{x}^a_{t_i}\}$$

$$\mathbf{z}^{\{a,+,-\}}_{d,t_i} := E_d(\mathbf{x}^{\{a,+,-\}}_{t_i}),$$
$$\mathbf{z}^{\{a,+,-\}}_{b,t_i} := E_b(\mathbf{x}^{\{a,+,-\}}_{t_i})$$

      2. Compute relative difficulty w

$$w\big((\mathbf{z}^a_{d,t_i}, \mathbf{z}^-_{d,t_i}), (\mathbf{z}^a_{b,t_i}, \mathbf{z}^-_{b,t_i})\big) = 1 + \frac{\tilde{D}(\mathbf{z}^a_{d,t_i}, \mathbf{z}^-_{d,t_i})}{\tilde{D}(\mathbf{z}^a_{d,t_i}, \mathbf{z}^-_{d,t_i}) + \tilde{D}(\mathbf{z}^a_{b,t_i}, \mathbf{z}^-_{b,t_i})},$$

      where $\tilde{D}(\cdot,\cdot)$ is the normalized distance among the batch $\mathcal{X}_t$
      3. Update E_b and E_d (with relative difficulty w from Step 2) by updating θ_b and θ_d as

$$\theta_b \leftarrow \theta_b - \alpha \nabla_{\theta_b} \frac{1}{|\mathcal{X}_t|} \sum_{\mathcal{X}_t} \hat{\mathcal{L}}_{tri}(\mathbf{z}^a_d, \mathbf{z}^+_d, \mathbf{z}^-_d)$$

$$\theta_d \leftarrow \theta_d - \alpha \nabla_{\theta_d} \frac{1}{|\mathcal{X}_t|} \sum_{\mathcal{X}_t} \hat{\mathcal{L}}_{w-ctr}(\mathbf{z}^a_b, \mathbf{z}^+_b, \mathbf{z}^-_b)$$

    end for
  end for
end while

so that two embeddings can be distant apart even though both are similar to the anchor point. If we are given τ^+ (as shown in (1)) in advance or use it as a hyperparameter, we can further mitigate wrongly selected negative samples and our final loss is as

$$\mathcal{L}_{wcl} = \mathbb{E}\left[-\log \frac{e^{\mathbf{z}^{a\intercal}_d \mathbf{z}^+_d}}{e^{\mathbf{z}^{a\intercal}_d \mathbf{z}^+_d} + g'} \right],$$

where $g' = \frac{1}{\tau^-}\left(w(\mathbf{z}^a, \mathbf{z}^-_b, \mathbf{z}^-_d) \cdot e^{\mathbf{z}^{a\intercal}_d \mathbf{z}^-_d} - \tau^+ e^{\mathbf{z}^{a\intercal}_d \mathbf{z}^+_d}\right)$. Then our final objective is to minimize the loss *w.r.t.* both θ_d and θ_b as

$$\arg\min_{\theta_d, \theta_b} \Big\{ \mathcal{L} := \mathcal{L}_{tri} + \mathcal{L}_{wcl} \Big\}.$$

The overview of the proposed representation learning method is illustrated in Figure 2. As in the figure, negative samples have different weights based on the relative difficulty, and hard negatives have larger weights, *i.e.*, $w_{hard} \gg w_{easy}$. Algorithm 1 describes the updating scheme of our method. We denote $\hat{\mathcal{L}}$ as an empirical version of each loss.

6. Experiments

In this section, we evaluate the validity of the proposed method by comparing it with state-of-the-art methods of image classification and classification on fairness benchmarks.

6.1. Image Classification

We evaluate our model (WCL) with the related methods for debiased contrastive learning that are discussed in Section 4.2: DCL [8] and HCL [36]. We compare the methods on the famous image classification datasets: CIFAR-10, CIFAR-100 [23], CelebA [27], Waterbirds [37]. For CIFAR-10 and CIFAR-100, we use ResNet50 [13] as the backbone encoder structure, and ResNet18 for all other datasets. For our method, the biased encoder with triplet loss, E_b, also has the same structure with the same optimization scheme. For CelebA dataset, we consider attractiveness as target label by following the previous works [32, 34] in bias mitigation. Attractive males are known to have the worst accuracy compared to other groups. The Waterbirds dataset predicts whether the bird in the image is waterbird or landbird, and the background (water or land) is known to make bias.

All models are trained for 500 epochs and $N = 1$ as $M = 254$ and batch size is 128. By following the previous works [8, 36], we use $\tau^+ = 0.1$ for CIFAR-10, 0.05 for CIFAR-100, and 0.3 for CelebA and Waterbirds dataset, respectively. We use $\beta = 1$ for HCL. We also set temperature $t = 0.5$ for all methods. For the downstream classification task, we conducted linear probing trained on the output of the penultimate layer of each model. We conducted all experiments on AMD 3860X CPU and RTX3090 GPUs.

6.1.1 Quantitative Results

In Table 1, we show the results of test accuracy on CIFAR-10 and CIFAR-100. ACC (worst) indicates the worst class accuracy, which is considered to be mostly biased against. Here, we include JTT [26] which aims to debias in the classification task. Compared with other methods, the classifier trained with the proposed debiased encoder E_d outperforms all accuracy measures in both datasets. It is interesting to note that it not only improved the worst group accuracy but also overall accuracy. This shows that our method has better generalization and robustness.

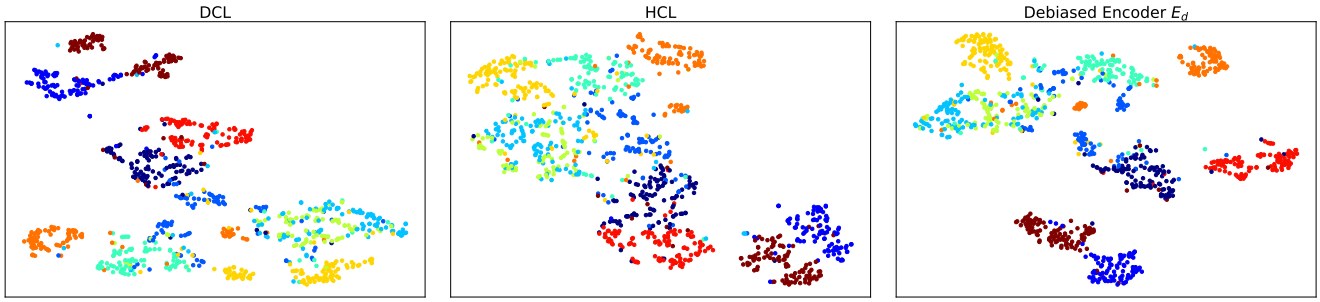
To validate the claim that we leverage relative difficulty by bias amplifying encoder, we achieve significantly lower

		CIFAR-10			CIFAR-100		
Method	Y	ACC (top-1)	ACC (top-5)	ACC (worst)	ACC (top-1)	ACC (top-5)	ACC (worst)
JTT [26]	O	85.67 ± 0.7	99.65 ± 0.2	72.33± 0.5	61.66 ± 0.6	83.53 ± 0.9	24.00 ± 1.5
SimCLR [4]	×	89.12 ± 0.6	99.74 ± 0.1	75.7 ± 0.4	64.86 ± 0.6	89.67 ± 0.3	20.00 ± 0.2
DCL [8]	×	91.66 ±0.3	99.78 ±0.1	81.2±0.2	68.26 ±0.3	91.19 ±0.1	20.00 ±0.2
HCL [36]	×	91.25 ±0.2	99.78 ±0.1	81.5 ±0.2	68.73 ±0.4	91.19 ±0.1	29.00 ±0.8
WCL (E_d)	×	**92.71±0.3**	**99.84±0.1**	**83.3±0.8**	**69.09±0.2**	**91.63±0.3**	**31.00±0.7**
WCL (E_b)	×	75.61 ±0.7	98.61 ± 0.4	52.6 ± 0.5	41.61± 0.3	69.26 ± 0.2	1.0 ± 0.5

Table 1. Performance evaluation on CIFAR-10 and CIFAR-100.

		Waterbirds [37]		CelebA [27]	
Method	Y	ACC (top-1)	ACC (worst)	ACC (top-1)	ACC (worst)
JTT [26]	O	77.81 ± 2.3	70.00 ± 1.5	76.83± 1.3	67.66 ± 0.5
SimCLR [4]	×	77.80 ± 1.5	0.00	78.61 ± 1.5	44.30 ± 0.7
DCL [8]	×	65.80 ± 1.7	4.51 ± 1.2	77.12 ± 1.6	44.95 ± 0.3
HCL [36]	×	69.31 ± 1.2	5.26 ± 1.1	76.13 ± 2.1	52.13 ± 0.8
WCL (E_d)	×	76.92 ± 0.3	**31.58 ± 3.5**	78.11 ± 2.3	**57.40 ± 1.2**
WCL (E_b)	×	73.64 ± 1.4	14.29 ± 1.5	58.84 ± 2.5	39.79 ± 1.3

Table 2. Performance evaluation on Waterbirds and CelebA dataset. Note thet JTT is supervised learning method. Among the self-supervised learning methods, WCL (ours) achieves the best worst group accuracy with comparable overall performance.



(a) DCL [8] (b) HCL [36] (c) WCL (Ours)

Figure 3. t-SNE visualization [40] of the learned representations on CIFAR-10. The colors indicate different classes. WCL (ours) achieves the best separation with small variance within a class.

accuracy, especially ACC (worst) when we train a classifier with biased encoder E_b. The gap between ACC (top-1) and ACC (worst) is dramatically larger than other methods. For example in CIFAR-100, for the encoder trained with triplet loss, E_b, it barely learns any information about the worst class. This supports our claim that triplet loss amplifies the bias and makes a model focus more on easy samples.

Moreover, we evaluated the methods on biased image benchmark datasets: CelebA and Waterbirds in Table 2. Compared to other contrastive learning methods, WCL significantly outperforms in terms of worst group accuracy. In particular, WCL could achieve almost 6 times better than the state-of-the-art (HCL) in CelebA dataset. In addition, we observed other debiasing methods (DCL, HCL) sacrifice overall accuracy to improve worst group accuracy. In contrast, we could substantially improve the worst group accuracy while maintaining the overall accuracy. The results suggest that the effectiveness of employing relative difficulty is more noticeable with the biased data. Also, the results show that the contrastive learning approaches achieve better generalization and robustness than JTT under multi-class scenarios by leveraging the generality of the learned representations [4] with data augmentation. Additional experiments on ImageNet-100 and weight analysis can be found in Appendix.

6.1.2 Qualitative Results

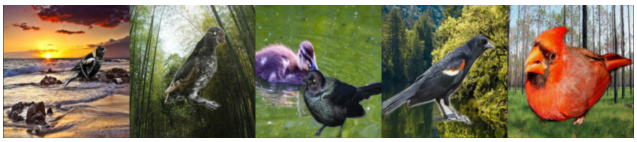
In Figure 3, we illustrate t-SNE visualization [40] of the learned representation by different methods. The colors of the points indicate 10 different classes in CIFAR-10 dataset. The difficulty-based reweighing method (WCL, E_d) achieves better separation between the class than others. Also, it is also interesting to note that the sample distributions within one class have a high variance for other methods than ours. This can lead to overlapping between distributions and cause biased predictions. In contrast, ours yields a relatively small variance, which means it handled

the hard samples well.

6.1.3 Sanity Check of Relative Distance

To ensure the relative distance, $w(\cdot)$, represents the biases, we depict top-5 easy and hard negatives from a randomly selected anchor in Figure 4. Given an anchor (waterbird on water), we observe three *landbird on land* samples in easy negatives, while having two *landbird on water* samples in hard negatives. Since the background of the image has spurious correlation [26] to the label, we observe that the proposed framework correctly realizes the spurious correlation and penalizes them while training E_d.

(a) Top-5 easy negatives



(b) Top-5 hard negatives

Figure 4. Visualization of top-5 easy and hard negative samples for an anchor (waterbird on water). Easy negative includes three landbird on land samples, while hard negative contains two landbird on water samples.

6.2. Classification in Fairness Benchmarks

Ethical problem [1, 6, 16, 21, 46] of algorithmic decision making is getting over the horizon as it is widely used in practice, such as risk assessment [9], credit limit assignment [41], *etc*. With this fairness perspective, the bias-conflicting samples are usually concentrated on certain groups, *i.e.,* unprivileged group. Subsequently, unfavorable predictions are more likely to occur, *i.e.,* higher false negative rates, on the unprivileged group, *e.g.,* black, female, *etc*.

To address this, Hardt *et al.* [11] post-process the outcome of a biased classifier to achieve equalized odds (EOd), $P(\hat{Y} = Y|A = 0) = P(\hat{Y} = Y|A = 1)$, where $\hat{Y}$, Y, and A are predicted label, target label, and sensitive attribute (*e.g.,* race and gender), respectively. To learn a fair representation, Madras *et al.* [29] employs adversarial learning strategy [**?**] to learn representation independent from sensitive attributes. However, these methods all require sensitive information in hand to achieve group fairness.

In contrast, our method takes Rawlsian Max-Min fairness [35] approach to address biased representation learning, which is to minimize the maximum sample loss, and this aligns with the motivation of this paper: focusing on hard samples for debiasing. Learning from Failure (LfF) [30] also takes this approach, but it only applies to supervised tasks as they use GCE loss. On the other hand, our method learns fair representation in a self-supervised manner by debiasing hard negatives of contrastive learning.

We evaluate our method on two fairness benchmark datasets. Adult [22] dataset has 14 features with *gender* as the sensitive attribute and is to predict if the annual income exceeds 50K. COMPAS [9] is to predict if each person gets rearrested within two years, where the sensitive attribute is *race*. We use MLP with ReLU activation followed by batch normalization for all contrastive learning methods and LfF. By following the original setup of LfF, we find $q \in [0, 1]$ with grid search that achieves the best validation accuracy. We randomly split each dataset into train, validation, and test sets with 70%, 15%, 15% ratio with 5 repetitions.

In Table 3, we demonstrate the performance and fairness violation in binary classification. The results validate that contrastive learning with difficulty referencing outperforms the existing methods [8, 36]. Also, note that even without a sensitive attribute, we could achieve comparable fairness with LAFTR [29] which requires a sensitive attribute.

			Adult [22]		COMPAS [9]	
Method	A	Y	ACC	EOd	ACC	EOd
Baseline	×	O	84.3	0.114	65.2	0.289
Eq. Odds [11]	O	O	82.8	0.055	61.8	0.115
LAFTR [29]	O	O	83.9	0.083	62.6	0.098
LfF [30]	×	O	82.1	0.185	62.5	0.124
DCL [8]	×	×	81.8	0.136	61.9	0.195
HCL [36]	×	×	81.9	0.132	62.0	0.198
WCL (E_d, ours)	×	×	82.6	0.104	64.3	0.113

Table 3. Evaluation of the methods on performance and fairness on Adult and COMPAS dataset. A and Y indicate necessity of sensitive information and target label for each model. EOd measures fairness violation, which is lower the better. WCL achieves comparable or better results without A and Y information.

7. Conclusion and Future Directions

We propose a novel contrastive learning strategy to highlight hard negative samples in data bias perspective. This takes Rawlsian Max-Min fairness [35] approach to address biased representation learning. We propose triplet loss to train a bias-amplifying model and leverage the relative distance surrogate to obtain the relative difficulty, which is grounded on the theoretical finding of triplet loss. Such relative difficulty leads to hard negative sampling so that we can train the debiased model in an unsupervised manner.

Acknowledgements This work was partially supported by the EMBRIO Institute, contract #2120200, a National Science Foundation (NSF) Biology Integration Institute, Purdue's Elmore ECE Emerging Frontiers Center, and NSF IIS #1955890, IIS #2146091.

References

[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *AAAI*, volume 33, pages 2412–2420, 2019. 8

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, pages 528–539, 2020. 1, 2

[3] Junyi Chai and Xiaoqian Wang. Self-supervised fair representation learning without demographics. In *NeurIPS*, 2022. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1, 3, 7

[5] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005. 1

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. 8

[7] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021. 1

[8] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *arXiv preprint arXiv:2007.00224*, 2020. 1, 2, 3, 6, 7, 8

[9] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.*, 4(1):eaao5580, 2018. 8

[10] 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. 2

[11] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016. 8

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[14] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, pages 4182–4192, 2020. 2

[15] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *WACV*, pages 2785–2795, 2022. 2

[16] Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. In *AAAI*, volume 36, pages 6988–6995, 2022. 8

[17] Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *AAAI*, volume 35, pages 7908–7916, 2021. 1

[18] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *arXiv preprint arXiv:2010.01028*, 2020. 2

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, volume 33, pages 18661–18673, 2020. 1, 2, 4

[20] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, pages 9012–9020, 2019. 2

[21] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. Model-agnostic characterization of fairness trade-offs. *arXiv preprint arXiv:2004.03424*, 2020. 8

[22] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996. 8

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[24] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*, volume 33, pages 728–740, 2020. 2

[25] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *NeurIPS*, volume 34, 2021. 1, 2, 4

[26] 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 *ICML*, pages 6781–6792, 2021. 2, 6, 7, 8

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 6, 7

[28] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018. 1

[29] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018. 8

[30] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *NeurIPS*, volume 33, pages 20673–20684, 2020. 2, 4, 8

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

[32] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *AAAI*, volume 35, pages 2403–2411, 2021. 6

[33] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. *arXiv preprint arXiv:2203.16209*, 2022. 2

[34] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *CVPR*, pages 8227–8236, 2019. 6

[35] John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001. 8

[36] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 1, 2, 4, 6, 7, 8

[37] 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. 6, 7

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2, 4

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794. Springer, 2020. 1

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 7

[41] Neil Vigdor. Apple card investigated after gender discrimination complaints. *The New York Times*, 2019. 8

[42] 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. 2

[43] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 1, 2

[44] Qiying Yu, Jieming Lou, Xianyuan Zhan, Qizhang Li, Wangmeng Zuo, Yang Liu, and Jingjing Liu. Adversarial contrastive learning via asymmetric infonce. In *ECCV*, pages 53–69. Springer, 2022. 2

[45] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, volume 31, 2018. 4

[46] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In *NeurIPS*, pages 15675–15685, 2019. 8