# Fighting Fire with Fire: Contrastive Debiasing without Bias-free Data via Generative Bias-transformation

Yeonsung Jung[1]   Hajin Shim[1]   June Yong Yang[1]   Eunho Yang[1,2]

## Abstract

Deep neural networks (DNNs), despite their impressive ability to generalize over-capacity networks, often rely heavily on malignant bias as shortcuts instead of task-related information for discriminative tasks. To address this problem, recent studies utilize auxiliary information related to the bias, which is rarely obtainable in practice, or sift through a handful of bias-free samples for debiasing. However, the success of these methods is not always guaranteed due to the unfulfilled presumptions. In this paper, we propose a novel method, Contrastive Debiasing via Generative Bias-transformation (CDvG), which works without explicit bias labels or bias-free samples. Motivated by our observation that not only discriminative models but also image translation models tend to focus on the malignant bias, CDvG employs an image translation model to transform one bias mode into another while preserving the task-relevant information. Additionally, the bias-transformed views are set against each other through contrastive learning to learn bias-invariant representations. Our method demonstrates superior performance compared to prior approaches, especially when bias-free samples are scarce or absent. Furthermore, CDvG can be integrated with the methods that focus on bias-free samples in a plug-and-play manner for additional enhancements, as demonstrated by diverse experimental results.

## 1. Introduction

Recent advances in deep learning have showcased that DNNs are capable of reaching state-of-the-art performance in various fields of machine learning, such as computer vision (He et al., 2016), natural language processing (Brown et al., 2020), reinforcement learning (Mnih et al., 2016) and more. However, it is also known that the over-parameterized nature of DNNs not only exposes them to general overfitting but also renders them susceptible to biases present in collected datasets (Torralba & Efros, 2011), which are detrimental to the generalizability. In supervised learning, neural networks tend to prefer shortcut solutions based on biases rather than real signals (Zhu et al., 2017b; Li et al., 2018). These spurious correlations do not provide task-related information, and DNNs that use these malignant biases, which are easier to perceive compared to the signal, will ultimately fail on future data. For instance, a classifier trained to identify *car racing* images using a dataset dominated by the track will exploit the track road information. However, the classifier will fail to perform well on images of off-road rallies. To this end, debiasing is imperative in utilizing DNNs for real-world applications.

A tautological solution to the bias problem is to construct a bias-free dataset from the start. However, curating a dataset devoid of all bias is extremely costly at best and generally infeasible. A more practical attempt at neutralizing dataset bias is to fortify a dataset with explicit supervision with regard to the bias (Kim et al., 2019; Sagawa* et al., 2020). However, additional expenditure of human labor in procuring such information cannot be avoided, which renders the option less appealing.

In most cases where such explicit supervision for bias is absent, the following two lines of work are recently proposed. One line of work mitigates the influence of bias by leveraging the bias type (e.g. texture) (Bahng et al., 2020; Geirhos et al., 2019; Wang et al., 2019; Hong & Yang, 2021) to design bias-oriented auxiliary models or to augment texture-perturbed samples. However, such prior knowledge of bias is by no means guaranteed, and even with such information, designing bias-oriented architectures is not always straightforward. The other line of work focuses on bias-free samples by leveraging the knowledge that malignant biases are learned faster than task-relevant features (Nam et al., 2020; Lee et al., 2021; Kim et al., 2021). However, these methods tend to break down in regimes where bias-free samples are scarce or absent (See Section 3.1).

[1]Graduate School of AI, Korea Advanced Institute of Science and Technology (KAIST), Seoul, South Korea [2]AITRICS, Seoul, South Korea. Correspondence to: Yeonsung Jung <ys.jung@kaist.ac.kr>, Eunho Yang <eunhoy@kaist.ac.kr>.

(a) Colored MNIST



(b) Corrupted CIFAR-10



(c) BAR



(d) BFFHQ
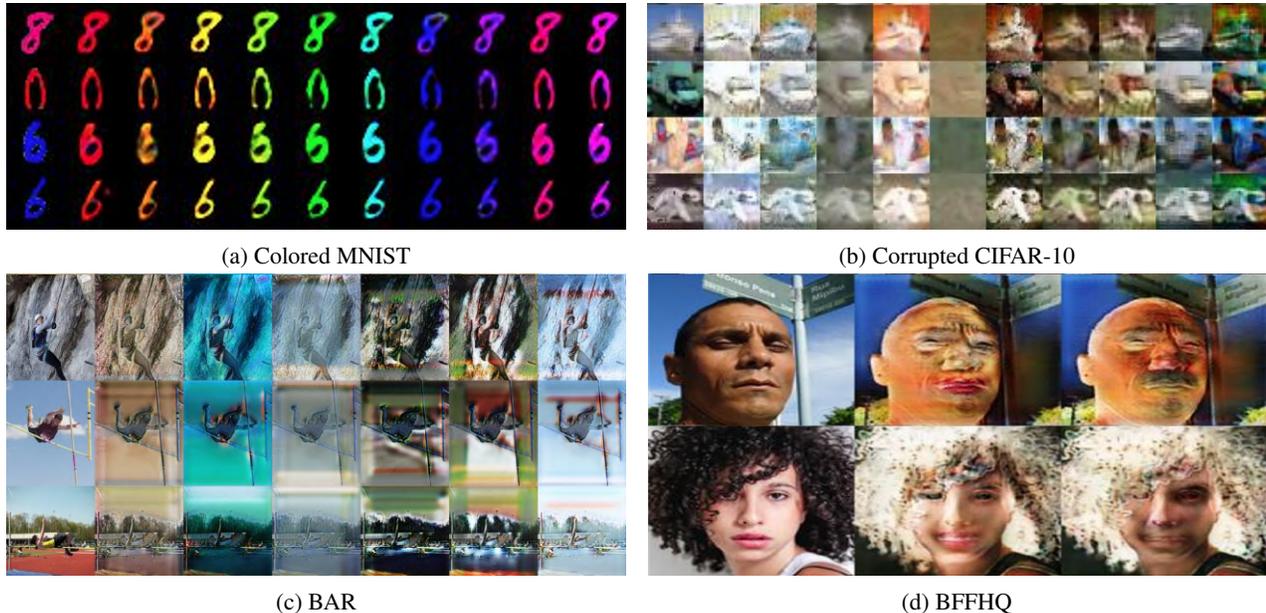
Figure 1: Bias-transformed images on synthetic and real-world biased datasets. The leftmost column contains the original images and each of remaining columns is the bias-transformed image of each target domain. The resulting images show the translated bias attributes such as color (Colored MNIST), texture (Corrupted CIFAR-10), background (BAR), and gender (BFFHQ), respectively. (For details about the datasets, we refer the reader to Appendix A.)

To tackle these shortcomings, we propose a general debiasing method, Contrastive Debiasing via Generative Bias-transformation (CDvG), that can operate even in regimes devoid of bias-free data. CDvG contrasts the bias modes within the dataset against each other to attenuate the influence of bias while effectively learning the task-relevant information contained in all samples.

Through preliminary experiments, we find that image-to-image translation models based on Generative Adversarial Networks (GANs) favor learning malignant biases over task-relevant signals, as discriminative models are known to do (See Section 2.2). Motivated by the observation, we train a biased image translation model without bias labels that learns the bias distribution over the signal, allowing us to transform the bias feature of a given input to another bias. Using the trained bias-translation model, we synthesize novel views with altered bias features.

Fighting fire with fire, we pit one bias type against another via contrastive learning. By maximizing agreement between the views with different biases, the model is encouraged to learn bias-invariant representations. Unlike existing methods, CDvG does not require explicit supervision, domain knowledge, or other meta-knowledge, such as the existence of bias-free samples.

Our contributions are three-fold:

- We experimentally observe that image-to-image trans-

lation models that utilize a domain classifier also favor malignant biases as discriminative models do (Sec. 2).

- We propose a novel approach for addressing the highly biased issue without the need for bias labels, bias type information, or even the existence of bias-free samples by employing an image translation model and contrastive learning (Sec. 3).

- Our method can be integrated with the methods that focus on bias-free samples in a plug-and-play manner for further enhancement. Extensive experiments on diverse datasets empirically demonstrate the effectiveness of our method, especially when bias-free samples are extremely scarce or absent (Sec. 4).

## 2. Image Translation Models for Generating Bias-Free Samples

Previous approaches to addressing the bias problem are to either deploy manpower or algorithms to salvage a sufficient amount of bias-free samples from the contaminated dataset. However, under circumstances where bias-free samples are almost nonexistent, there are not enough bias-free samples to begin with. Instead of salvaging, we first opt to synthesize bias-free samples using image translation models. In other words, we aim to translate samples aligned with the bias to bias-free samples by changing a given bias factor to another. In this section, we first define our setup in Section 2.1 and
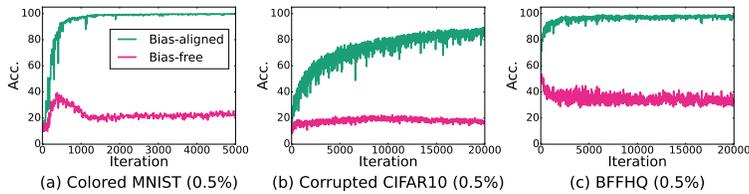
Figure 2: Unbiased accuracy during StarGAN training.
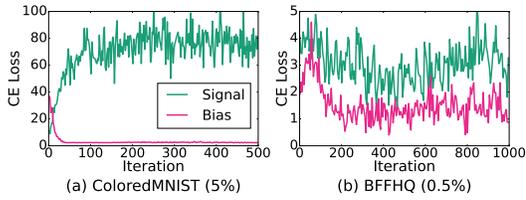


Figure 3: CE loss during StarGAN training.

describe our observations of how translation models behave on biased datasets, which is a key part of our algorithm, in Section 2.2.

## 2.1. Setup

To formally define our target task, we introduce the following random variables: input image $X$, target label $Y$, signal $S$, bias $B$ and other attributes $O$ comprising input $X = (S, B, O)$. Here, $B$ is a feature unrelated to $Y$, that is, $Y$ and $B$ are independent given $S$. We further define the random variable for bias label $Y_B$ which is hidden.

Ideally, the image classification model parameterized by $\theta$ predicts the label based on the signal as $P_\theta(Y|X) = P_\theta(Y|B, S) = P_\theta(Y|S)$. However, when the training data consists of highly but spuriously correlated bias and target, i.e., $P_\theta(Y^{train}|B^{train}) \approx 1$, predicting $Y$ based on $B$ is also one of the possible solutions that can be deemed effective in the training phase. In this paper, when the sample $x = (s, b, o)$ consists of a correlated signal and a bias (i.e. high $P_\theta(b|s)$), $x$ is called *bias-aligned* and the opposite case is called *bias-free*. We tackle the case where $B$ is easier to perceive than $S$ so that the spurious correlation between $B$ and $Y$ is *malignant* in that the model preferentially takes $B$ as a clue to predict $Y$ over $S$ (Nam et al., 2020). This is obviously an unintended consequence and impairs generalizability due to the discrepancy of the bias-target correlation $Y|B$ between the training and test phases.

## 2.2. Behavior of Image Translation Models Under Bias

We first examine whether GAN-based image translation models (Isola et al., 2017; Zhu et al., 2017a; Choi et al., 2018; 2020), which render an image $x$ from a source domain $y$ to a target domain $y'$, are capable of generating bias-free samples. Ideally, they find out the representative characteristics of the target domain $y'$ and combine them with the input image.

Given that discriminative models have been shown to be susceptible to bias, it is not far-fetched to first enquire whether generative models also carry the same frailty. Specifically, it is intuitive to suspect that the family of GANs may be prone to bias due to the presence of a discriminator component, which is known to be bias-pregnable.

In addition, Zhu et al. (2017a), a representative milestone of image translation, presented a number of typical failure modes that when the source domain is an apple and the target domain is an orange, a transformed image is not the orange counterpart of the input, but an apple with the color and texture of an orange. This implies that the image translation model perceives color and texture rather than shape as the representative traits for the target domain, even without using a highly biased dataset. With this in consideration, it is plausible to speculate that this phenomenon would be exacerbated when handling highly biased datasets.

To verify the impact of bias on image translation models, we examine the behavior of StarGAN (Choi et al., 2018) on biased datasets. We observe that the translation model interacting with the biased domain discriminator is also prone to translate biases rather than task-related domain features to satisfy the domain classifier (See Figure 1). The selection of StarGAN is driven by its cost-effectiveness relative to recent models and its status as a well-established, extensively studied in the field. StarGAN's overall size is comparable to that of auxiliary models, such as ResNet18 (He et al., 2016), commonly used in recent debiasing approaches. This makes it a cost-effective alternative to existing baseline models. Furthermore, we found that recent image translation models, which use a multi-task discriminator, are also inherently prone to malignant bias when using a domain classifier. In Appendix E, we empirically show that replacing the multi-task discriminator with a domain classifier in StarGANv2 (Choi et al., 2020) leads to a pronounced focus on malignant bias, suggesting that this finding may extend to other recent models.

**Discriminators of translation models are also prone to be biased.** StarGAN introduces an auxiliary domain classifier $D_{cls}$ on top of discriminator $D$ to enable translation between multiple domains. The domain classifier $D_{cls}$, trained on the real image with domain labels, learns to classify images with representative traits of the domains by optimizing the domain classification loss of real images $\mathcal{L}_{cls}^r = \mathbb{E}_{(x,y)\sim\mathcal{D}}[-\log D_{cls}(y|x)]$.

However, on the biased dataset, we observed that $D_{cls}$ had absorbed bias attributes as representative traits of the domains during a training phase. To quantitatively evaluate whether $D_{cls}$ is biased, we measure the classification loss
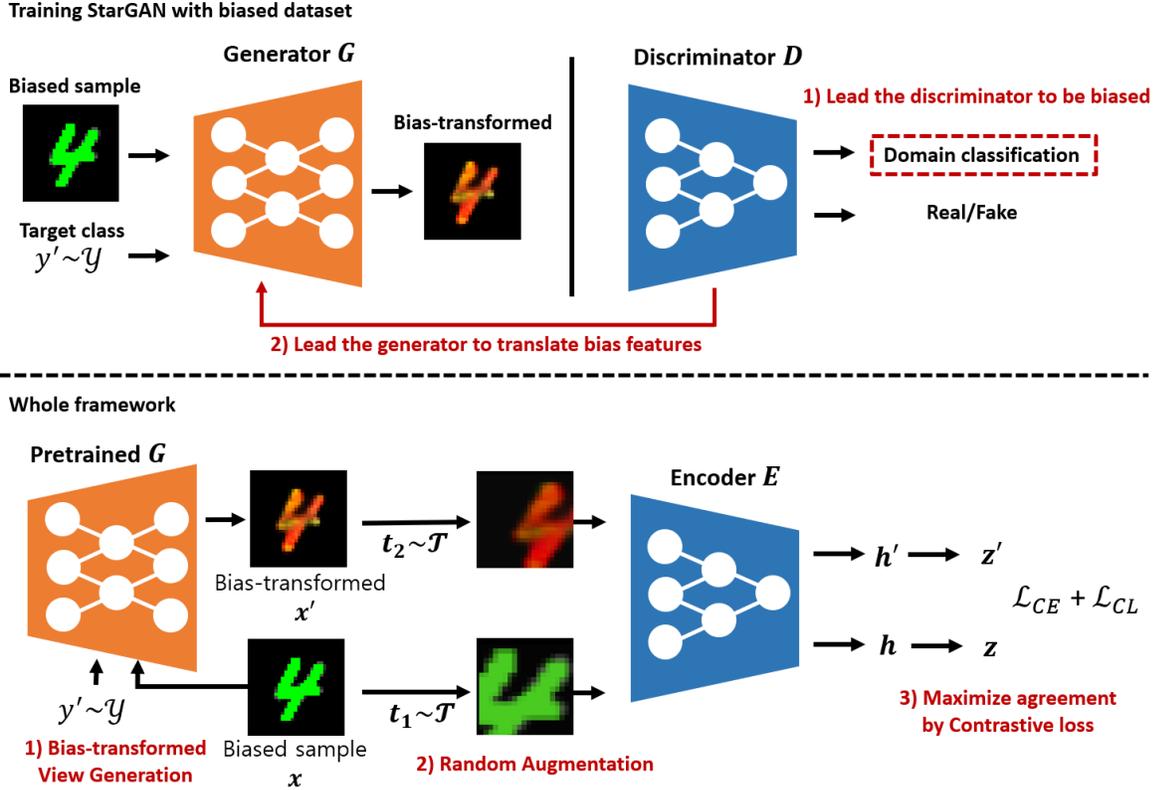
Figure 4: Illustration of our Contrastive Debiasing via Generative Bias-transformation (CDvG).

on the unbiased dataset with $D_{cls}$ while training StarGAN on the biased dataset Colored MNIST, Corrupted CIFAR10, and BFFHQ which have color bias, texture bias, and gender bias, respectively (See Figure 1). Figure 2 shows the classification losses of bias-aligned and bias-free samples. We observed that the accuracy of bias-aligned samples increases to nearly 100%, however, the accuracy of bias-free samples is low depending on how malignant the bias of each data set is. Therefore, we conclude that $D_{cls}$ utilizes the bias rather than the task-related features, as discriminative models are known to do so. Please notice that, given the similar architecture, identical biased trainset, and shared classification objective of the translation model's discriminator and the main task's discriminative model, it is natural to expect that both models would learn the same bias.

**A biased domain classifier induces a biased translation generator.** By optimizing the domain classification loss $\mathcal{L}_{cls}^f = \mathbb{E}_{(x,y)\sim\mathcal{D},y'\sim\mathcal{Y}}[-\log D_{cls}(y'|G(x,y'))]$ of translated fake images $G(x,y')$, biased $D_{cls}$ induces the generator $G$ to translate the bias attribute of image $x$ into the other bias correlated to the randomly sampled target domain $y'$ rather than the task-relevant signal. As a result, $G$ becomes a bias-translator. This phenomenon becomes more noticeable as the bias is malignant for $D_{cls}$ - that is, the more scarce the bias-free samples are or the easier the bias is to

perceive. To quantitatively evaluate whether the translation model truly favors learning biases over task-relevant signals, we measure the classification loss of translated images $x' = G(x,y')$ with the bias classifier $C_B$ and the signal classifier $C_S$ while training StarGAN on the biased dataset Colored MNIST and BFFHQ. As alternatives of oracles, the classifiers $C_B$ and $C_S$ are trained with the bias label $y_B$ and the true class label $y$, respectively, on the unbiased Colored MNIST and FFHQ. The bias and signal loss are defined as $\mathbb{E}_{(x,y)\sim\mathcal{D},y'\sim Cat(|\mathcal{Y}|)}[\mathcal{L}_{CE}(C_B(x'),y')]$ and $\mathbb{E}_{(x,y)\sim\mathcal{D},y'\sim Cat(|\mathcal{Y}|)}[\mathcal{L}_{CE}(C_S(x'),y')]$, respectively.

As expected, the results depicted in Figure 3 demonstrate that the bias is more favorable to the translation model. For Colored MNIST, the bias loss quickly declines to zero, but the signal loss increases rapidly since $G$ concentrates on the color biases rather than digits. In the case of BFFHQ, the bias is less malignant, resulting in a relatively small gap between signal and bias loss, but still places significant emphasis on the bias.

We also present the qualitative results in Figure 1. Interestingly, we found that the translated results retain their contents to some extent and color, texture, background (e.g. rock wall for climbing, water for fishing), and gender characteristics (e.g. makeup for female, beard for male), which are the respective bias features of the datasets, are altered.

4

Although the generated images of BFFHQ in Figure 1d appear dissimilar to real images when viewed by humans, the core contents remain discernible while the gender-specific features emerge as pink lips and eye makeup in the second column and thick eyebrows and beard in the third column. Furthermore, it is notable that the biased classifier judges that the bias-transformed images are well transformed into the target label with a high degree of confidence. These observations verify that there is a chance that image translation models trained on biased datasets can be employed to synthesize bias-free samples for debiasing.

## 3. Method

In this section, we empirically show that the recent debiasing methods break down when the assumption of the availability of bias-free samples is not met in Section 3.1. Acknowledging this limitation, we propose a novel debiasing method, referred to as CDvG, in Section 3.2 leveraging our empirical findings (Section 2.2) that image translation models are prone to be biased.

### 3.1. Learning without Bias-free Samples

Although recent debiasing methods (Bahng et al., 2020; Hong & Yang, 2021; Nam et al., 2020; Lee et al., 2021; Kim et al., 2021) work as intended when they can exploit the bias-free samples, they neglect the case where such bias-free samples are extremely scarce or absent. Thus, in regimes where a presumption that enough bias-free samples exist does not hold, their behaviors are practically unknown and are likely to break down. More specifically, in Table 1, we empirically demonstrate that previous methods fail to effectively debias the model (showing low performance, almost comparable to Vanilla) when the proportion of bias-free samples is 0%.

Furthermore, we investigate the sample reweighting scheme (Nam et al., 2020; Lee et al., 2021) that aims to balance bias-free and bias-aligned samples. We observed that as the proportion of bias-free samples decreases, the debiasing effect also decreases, due to the inadequate emphasis placed on bias-free samples. The detailed descriptions and results are deferred to Section C. Based on these observations, we consider a more general task of learning-under-bias that handles the situations where there are no to only a few bias-free samples as well as the cases where such samples are sufficiently provided.

### 3.2. Contrastive Debiasing without Bias-free Data via Bias-transformed Views

Motivated by the findings of Section 2, we propose a novel debiasing method referred to as Contrastive Debiasing via Generative Bias-transformation (CDvG). The method em-

---

**Algorithm 1** Contrastive Debiasing via Generative Bias-transformation (CDvG)

---
1: **Input:** Encoder $E$, Projection head $H$, Classifier $C$, Biased generator $G$, Augmentation family $\mathcal{T}$
2: **Data:** Training set $\mathcal{D} = \{(x, y)\} \subset \mathcal{X} \times \mathcal{Y}$
3: **for** minibatch $\{(x_k, y_k)\}_{k=1}^{N}$ **do**
4:     $y_k' \sim Categorical(|\mathcal{Y}|)$
5:     $x_k' = G(x_k, y_k')$
6:
7:     $t_1 \sim \mathcal{T}, t_2 \sim \mathcal{T}$
8:     $\tilde{x}_k, \tilde{x}_k' = t_1(x_k), t_2(x_k')$
9:
10:     **Update** $E, C, H$ to minimize
11:     $\sum_k (\mathcal{L}_{CE}(C(E(\tilde{x}_k)), y_k) + \mathcal{L}_{CE}(C(E(\tilde{x}_k')), y_k)) + \mathcal{L}_{CL}(E, H)$
12: **end for**

---

ploys the biased translation model to transform an image to have different biases corresponding to other classes and integrates it with the contrastive learning framework. The contrastive loss is employed to encourage the learning of bias-invariant representations without the need for explicitly identifying bias-free data. The bias-transformed views are contrasted against each other to attenuate the bias while effectively learning the features that are relevant to the task. The whole process is outlined in the following paragraphs and summarized in Algorithm 1.

First, we train StarGAN on a given biased training dataset to obtain the bias-transformation generator $G$. The bias-transformation generator $G$ is then utilized to generate diverse bias-transformed views $x' = G(x, y')$ of an input image $x$ with a target label $y'$, where $y'$ is uniformly sampled for every iteration (Step 1 of the whole framework in Figure 4, Line 4 and 5 in Alg. 1). Bias-translation with $G$ is only applied during training, not in the testing phase, due to the distribution shifts.

After the bias-transformation step, we additionally apply the random augmentation operators $t_1$ and $t_2$ to the original and the bias-transformed images respectively as $\tilde{x} = t_1(x)$, $\tilde{x}' = t_2(x')$, where $t_1$ and $t_2$ are sampled from the same augmentation family $\mathcal{T}$ (Step 2 of the whole framework in Figure 4, Line 7 and 8 in Alg. 1). According to Chen et al. (2020a), strong augmentations are essential to the performance of the contrastive learning framework as they prevent the encoder from easily finding a clue about the fact that two views come from the same image. By following Chen et al. (2020a), $\mathcal{T}$ is composed of the following sequential augmentations: random resized cropping and random horizontal flipping. Please note that we did not adopt color distortion methods because most biases are related to color.

Finally, the resulting learning objective is given by the com-

Table 1: The average and the standard deviation of accuracy over 3 runs. Ratio(%) represents that the proportion of bias-free samples. ✓ indicates that the model exploits auxiliary information of the bias and ✗ does not.

| Dataset | Ratio(%) | Vanilla ✗ | ReBias ✓ | LfF ✗ | DisEnt ✗ | BiaSwap ✗ | BPA ✗ | CDvG ✗ | CDvG+LfF ✗ |
|---|---|---|---|---|---|---|---|---|---|
| Colored MNIST | 0.0 | $12.53_{\pm0.92}$ | $14.64_{\pm0.50}$ | $13.16_{\pm1.87}$ | $11.65_{\pm0.61}$ | - | $10.08_{\pm0.04}$ | $\underline{96.16}_{\pm0.48}$ | $\mathbf{96.48}_{\pm0.19}$ |
| | 0.5 | $39.12_{\pm0.91}$ | $70.47_{\pm1.84}$ | $66.54_{\pm3.80}$ | $62.13_{\pm3.96}$ | 85.76 | $54.52_{\pm3.39}$ | $\underline{95.73}_{\pm0.14}$ | $\mathbf{96.20}_{\pm0.12}$ |
| | 1.0 | $56.02_{\pm1.81}$ | $87.40_{\pm0.78}$ | $79.83_{\pm2.23}$ | $75.49_{\pm0.21}$ | 83.74 | $72.63_{\pm0.27}$ | $\underline{96.13}_{\pm0.48}$ | $\mathbf{96.45}_{\pm0.19}$ |
| | 2.0 | $69.32_{\pm0.22}$ | $92.91_{\pm0.15}$ | $82.66_{\pm0.39}$ | $80.08_{\pm0.45}$ | 85.29 | $78.52_{\pm0.59}$ | $\underline{96.90}_{\pm0.22}$ | $\mathbf{96.97}_{\pm0.17}$ |
| | 5.0 | $83.93_{\pm0.89}$ | $\underline{96.96}_{\pm0.04}$ | $83.30_{\pm1.23}$ | $85.00_{\pm1.17}$ | 90.85 | $85.30_{\pm0.93}$ | $96.73_{\pm0.05}$ | $\mathbf{96.98}_{\pm0.05}$ |
| Corrupted CIFAR-10 | 0.0 | $16.05_{\pm0.13}$ | $21.93_{\pm0.37}$ | $15.88_{\pm0.45}$ | $18.76_{\pm0.88}$ | - | $17.14_{\pm1.54}$ | $\underline{28.44}_{\pm0.21}$ | $\mathbf{29.24}_{\pm0.47}$ |
| | 0.5 | $20.87_{\pm0.34}$ | $22.27_{\pm0.41}$ | $25.58_{\pm0.23}$ | $28.62_{\pm1.74}$ | 29.11 | $25.50_{\pm1.03}$ | $\underline{31.50}_{\pm0.33}$ | $\mathbf{39.07}_{\pm0.27}$ |
| | 1.0 | $24.05_{\pm0.61}$ | $25.72_{\pm0.20}$ | $30.68_{\pm0.50}$ | $32.31_{\pm0.03}$ | 32.54 | $26.86_{\pm0.69}$ | $\underline{33.25}_{\pm0.20}$ | $\mathbf{43.81}_{\pm0.35}$ |
| | 2.0 | $29.47_{\pm0.20}$ | $31.66_{\pm0.43}$ | $\underline{37.96}_{\pm1.09}$ | $36.51_{\pm2.34}$ | 35.25 | $27.47_{\pm1.46}$ | $35.16_{\pm0.23}$ | $\mathbf{47.45}_{\pm0.07}$ |
| | 5.0 | $41.12_{\pm0.16}$ | $43.43_{\pm0.41}$ | $\underline{48.49}_{\pm0.16}$ | $46.41_{\pm0.62}$ | 41.62 | $34.29_{\pm2.20}$ | $42.75_{\pm0.19}$ | $\mathbf{52.31}_{\pm0.13}$ |
| BFFHQ | 0.0 | $37.93_{\pm0.96}$ | $43.47_{\pm0.74}$ | $39.67_{\pm1.00}$ | $38.13_{\pm2.13}$ | - | $48.20_{\pm1.40}$ | $\underline{48.80}_{\pm1.18}$ | $\mathbf{49.60}_{\pm1.18}$ |
| | 0.5 | $52.40_{\pm1.88}$ | $56.80_{\pm1.56}$ | $\underline{58.07}_{\pm0.82}$ | $54.33_{\pm0.92}$ | - | $51.40_{\pm2.98}$ | $56.80_{\pm0.33}$ | $\mathbf{62.20}_{\pm0.45}$ |

bination of the cross entropy loss and the contrastive loss (Step 3 of the whole framework in Figure 4, Line 11 in Alg. 1). First, the encoder $E$ and the following classifier $C$ are optimized to minimize the cross entropy loss $L_{CE}$ which is applied to both $\tilde{x}$ and $\tilde{x}'$:

$$\min_{E,C} \mathcal{L}_{CE}(C(E(\tilde{x})), y) + \mathcal{L}_{CE}(C(E(\tilde{x}')), y).$$

Also, we train the encoder $E$ and the projection head $H$ to minimize the contrastive loss $\mathcal{L}_{CL}$ to tie the original view $\tilde{x}$ and the bias-transformed view $\tilde{x}'$ as $\mathcal{L}_{CL}(E, H) = \sum_{k=1}^{N} \ell(2k-1, 2k) + \ell(2k, 2k-1)$. The loss $\ell(i, j)$ for a positive pair $(i, j)$ is defined as

$$\ell(i, j) = -\log \frac{\exp(sim_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(sim_{i,k}/\tau)},$$

Where $sim_{i,j} = z_i^\top z_j / (\|z_i\| \|z_j\|)$ is the cosine similarity, and $z_i = H(E(\tilde{x}_i))$ is a projected representation of $\tilde{x}_i$ with the base encoder $E$ followed by the projection head $H$. By tying the original view $\tilde{x}$ and the bias-transformed view $\tilde{x}'$ as a positive pair, the encoder attempts to attenuate biases while emphasizing the true signals shared by the views via maximizing the mutual information between their latent representations.

## 4. Experiments

To validate the effectiveness of CDvG compared to recent debiasing methods, we conduct image classification experiments on standard benchmark datasets for debiasing. We first outline the experimental setup, including the datasets and baselines in Section 4.1. Then, we present the main results, which consist of comprehensive comparisons between our method and the baselines across various datasets in Section 4.2, and provide detailed analysis of the learned bias-invariant representations by visualizations in Section 4.3. Additionally, we conducted ablation studies to demonstrate the contributions of each component of CDvG toward the performance improvements in Section 4.4. Also, detailed information regarding the datasets is provided in Appendix A, while implementation specifics are covered in Appendix B."

### 4.1. Experimental settings

**Dataset** We experiment on {Colored MNIST, Biased MNIST, Corrupted CIFAR-10} and {BFFHQ, ImageNet-9 (IN-9), Waterbirds} which are synthetic datasets injected with synthetic biases and real-world datasets with natural biases, respectively. By setting the proportion of bias-free samples to a range of 0-5%, we evaluate the performance considering the highly biased setting (0.5-5%) as following the convention and, moreover, the most challenging scenario (0%) where bias-free samples are absent. The datasets with 0% bias-free samples are constructed by excluding bias-free samples from the datasets with a ratio of 0.5%.

**Baselines** To evaluate the performance of CDvG, we conducted comprehensive comparisons with various categories of debiasing methods. Specifically, we compared CDvG against HEX (Wang et al., 2019) and Rebias (Bahng et al., 2020), which leverage domain knowledge to address bias, as well as LfF (Nam et al., 2020), which intentionally trains an auxiliary biased classifier to identify and utilize bias-free samples. Moreover, we assessed DisEnt (Lee et al., 2021) and BiaSwap (Kim et al., 2021), which build upon the bias-free sample selection scheme employed by LfF, aiming to

Table 2: ImageNet-9 dataset.

| Dataset | Test type | Vanilla | RUBi | LfF | Rebias | CDvG+LfF |
|---------|-----------|---------|------|-----|--------|----------|
| | Biased | $94.0_{\pm0.1}$ | $93.9_{\pm0.2}$ | $91.2_{\pm0.1}$ | $94.0_{\pm0.2}$ | $\mathbf{95.2}_{\pm0.1}$ |
| IN-9 | Unbiased | $92.7_{\pm0.2}$ | $92.5_{\pm0.2}$ | $89.6_{\pm0.3}$ | $92.7_{\pm0.2}$ | $\mathbf{94.5}_{\pm0.1}$ |
| | IN-a | $30.5_{\pm0.5}$ | $31.0_{\pm0.2}$ | $29.4_{\pm0.8}$ | $30.5_{\pm0.2}$ | $\mathbf{34.6}_{\pm0.4}$ |

Table 3: Waterbirds dataset.

| Dataset | Test type | ResNet-18 | | | | ResNet-50 | | |
|---------|-----------|-----------|-----|-----|----------|-----------|------|----------|
| | | Vanilla | LfF | BPA | CDvG+LfF | ERM | EIIL | CDvG+LfF |
| Waterbirds | Unbiased | 84.63 | 85.48 | **87.05** | 86.25 | **97.30** | 96.90 | 91.30 |
| | Worst-group | 62.39 | 68.02 | 71.39 | **74.92** | 60.30 | 78.70 | **84.80** |

incorporate other biases into the instances. Regarding BiaSwap, we relied on available performance directly from the corresponding paper while leaving blank spaces for the others due to the unavailability of the official code. Additionally, we examined BPA (Seo et al., 2022) and EIIL (Creager et al., 2021) for further comparison.

## 4.2. Main Results

We report the results on the standard benchmark datasets to validate the effectiveness of CDvG for debiasing when bias labels are not available. We observed that the CDvG demonstrated superior performance when bias is more malignant, specifically when bias-free samples are scarce(0.5%) or absent(0%). Furthermore, to leverage bias-free samples, we propose to combine CDvG and reweighting-based methods such as LfF (Nam et al., 2020) ('CDvG+LfF' in Table 1). The combined version is consistently superior to the baselines with a low standard deviation in Table 1. In more detail, in the training process, we trained CDvG by concurrently applying LfF's sample-wise weighting to CE loss for both original images and bias-translated images. Therefore, 'CDvG+LfF' allows us to utilize both bias-aligned and bias-free samples to debias by translating bias-aligned samples into bias-free samples (CDvG) and giving more weights to bias-free samples (LfF). Please note that CDvG can be easily integrated with other debiasing methods in a plug-and-play manner.

**Synthetic datasets**   In Table 1, we observe that the CDvG performs well on Colored MNIST and Corrupted CIFAR-10 when the bias-free samples are scarce or absent, i.e., when the biases are more malignant. CDvG+LfF, which is our

integrated method, outperforms the baselines on the overall ratio by a large margin.

**Real-world datasets**   We conducted experiments on BFFHQ, IN-9, and Waterbirds to evaluate the performance in a real-world setting. CDvG+LfF shows improved performance with small standard deviations on BFFHQ compared to state-of-the-art and also shows improved or comparable performance on IN-9 and Waterbirds in Table 2 and Table 3, respectively. For Waterbirds, we borrow the performance on ResNet-18 and ResNet-50 directly from BPA and EIIL, respectively.

**Multiple biases datasets**   To verify our method can handle multiple types of biases, we conducted experiments on Biased MNIST which has 'digit color', 'type of background texture', 'background color', 'co-occurring letters', and 'colors of the co-occurring letters' as biases. We confirmed that CDvG shows better performance than baselines when multiple biases exist in Table 4. Also, we confirm that the translation model uncovers all sources of biases according to their intensity. We borrow the performance of SD (Pezeshki et al., 2021), GroupDPO (Sagawa* et al., 2020), and PGI (Ahmed et al., 2021) on ResNet-18 from Shrestha et al. (2022).

In summary, our method can handle various types of synthetic and real-world biases and works well on a wide range of bias ratios, especially when bias-free samples are scarce and even absent compared with state-of-the-art methods. Overall consistent results empirically demonstrate that our method debias better than previous approaches focusing only on bias-free samples.

## 4.3. Visualizations

We explore the interpretability of the learned bias-invariant representations using t-SNE (Van der Maaten & Hinton, 2008) and GradCAM (Selvaraju et al., 2017) on Colored

Table 4: Biased MNIST dataset.

| Dataset | ERM | SD | UpWt | GroupDPO | PGI | CDvG |
|---------|-----|-----|------|----------|-----|------|
| BiasedMNIST | 36.8 | 37.1 | 37.7 | 19.2 | 48.6 | **49.48** |
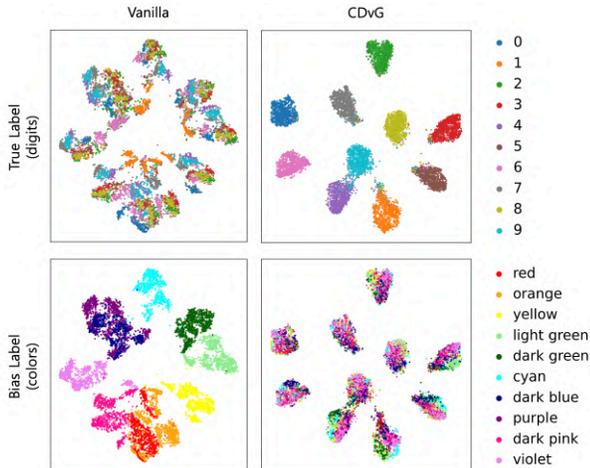
Figure 5: Visualization of learned representations of the vanilla model and CDvG using t-SNE in Colored MNIST.

MNIST and BFFHQ datasets, respectively. Our results on t-SNE show that the vanilla model heavily relies on the biased attribute 'color' for classification, as shown in the first column in Figure 5. In contrast, CDvG focuses more on the true label 'digits', as shown in the second column. This indicates that CDvG is better able to learn bias-invariant representations and is less dependent on biased attributes.

Similarly, in the case of GradCAM, we found that the vanilla model heavily relies on biased attributes 'gender' such as the beard of a male and the hair of a female, while CDvG is less dependent on these attributes and instead focuses more on facial features in Figure 6. These findings demonstrate that CDvG can learn more bias-invariant representations than the vanilla model.

### 4.4. Ablation study

We study the effectiveness of each component of CDvG framework. As shown in Table 5, CDvG without bias-transformation generator (CDvG w/o $G$), which is adding augmentation operator $\mathcal{T}$ and contrastive loss to vanilla, shows degradation in performance. This is due to the fact that the use of an $\mathcal{T}$ alone, without the bias-transformation generator, results in augmented samples that are still biased and ultimately contribute to the model becoming biased. However, when only the bias-transform generator $G$ is used for augmentation (CDvG w/o CL), there are 8.21%, 5.37%, 4.07%, 2.22%, and 1.43% improvements for each ratio, respectively. This demonstrates that the bias-transformation generator effectively augments bias-transformed images to some extent.

Despite this, it is important to note that the bias-transformation generator $G$ is not always a perfect model, as it may not fully translate biases due to the intensity of
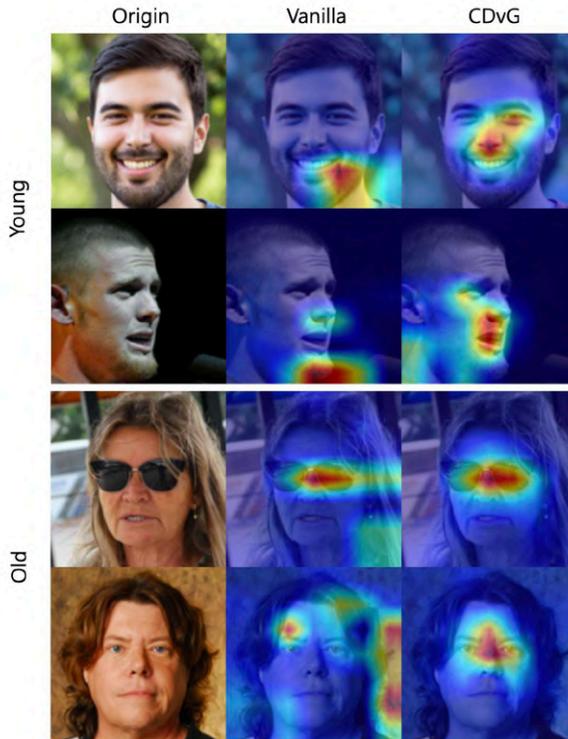


Figure 6: Visualization of Grad-CAM heatmaps for the labels 'Young' and 'Old' of BFFHQ using the vanilla model and the proposed CDvG.

the bias. Therefore, it is essential to induce bias-invariance and capture the true signal shared by the bias-transformed view and the original view by contrastive loss. The results of the ablation study show that the use of both the bias-transformation generator and contrastive loss (CDvG) results in the most significant improvement in performance, demonstrating that our whole framework effectively addresses the bias problem rather than a single component.

Table 5: Ablation study for bias-transformed view generation (w/o G) and for contrastive loss (w/o CL).

| Dataset | Ratio(%) | Vanilla | CDvG w/o $G$ | CDvG w/o CL | CDvG |
|---|---|---|---|---|---|
| Corrupted CIFAR-10 | 0.0 | $16.04_{\pm 0.13}$ | $18.29_{\pm 0.90}$ | $24.25_{\pm 0.15}$ | $\mathbf{28.44}_{\pm 0.21}$ |
| | 0.5 | $20.87_{\pm 0.34}$ | $20.75_{\pm 0.21}$ | $26.24_{\pm 0.12}$ | $\mathbf{31.50}_{\pm 0.33}$ |
| | 1.0 | $24.05_{\pm 0.61}$ | $21.37_{\pm 0.29}$ | $28.12_{\pm 0.21}$ | $\mathbf{33.25}_{\pm 0.20}$ |
| | 2.0 | $29.47_{\pm 0.20}$ | $24.50_{\pm 0.34}$ | $31.69_{\pm 0.41}$ | $\mathbf{35.16}_{\pm 0.23}$ |
| | 5.0 | $41.12_{\pm 0.16}$ | $31.13_{\pm 0.56}$ | $42.55_{\pm 0.50}$ | $\mathbf{42.75}_{\pm 0.19}$ |

## 5. Related Work

In real-world datasets across diverse domains, various kinds of incidental biases are strongly but spuriously correlated with task-related information. However, when the bias factors are more noticeable and easier to learn than the task-related signals, DNNs tend to lean on such biases (Nam

et al., 2020), causing failures in generalization. To counter this effect, several lines of work were developed.

**Explicit bias supervision** can be used to screen or mitigate the influence of bias. Kim et al. (2019) uses bias supervision to train an auxiliary network that helps in reducing the influence of bias. Sagawa* et al. (2020) uses bias supervision to group data samples for grouped distributionally robust optimization. Hong & Yang (2021) propose bias-contrastive loss and bias-balanced regression that encourages the model to pull together the samples in the same class with different bias features with balancing the target-bias distribution.

**Domain knowledge of biases.** When acquiring bias supervision is impractical, we can leverage domain-specific knowledge about bias type. For example, it was shown that ImageNet-trained classifiers exploit texture information in the image rather than information contained in the object of interest (.Gatys et al., 2017; Brendel & Bethge, 2019). Utilizing this fact, Geirhos et al. (2019) construct an augmented dataset by applying various textures to the images for texture debiasing. Also, Wang et al. (2019) and Bahng et al. (2020) exploit an auxiliary model which is carefully designed to capture biases over signals. However, these methods are grounded on obtaining such knowledge - which is often costly, or even impossible. In addition, designing a bias-oriented auxiliary model may not be so intuitive, depending on the type of bias.

**Without domain knowledge.** Using the fact that malignant biases are learned faster than salient task-relevant features, Nam et al. (2020) train an auxiliary model using generalized cross entropy (Zhang & Sabuncu, 2018) to emphasize malignant bias and assign more weight to the bias-free samples for the debiased follower. Seo et al. (2022) cluster the training samples and reweight the loss of each sample in a batch according to the assigned clusters. Lee et al. (2021) proposed feature-level data augmentation that disentangles bias features by using a biased auxiliary model obtained by following Nam et al. (2020) and swaps latent bias features. Kim et al. (2021) employs task-related features in abundant bias-aligned samples by synthesizing a new image that takes the bias-irrelevant core features from the biased sample and the bias attribute from the bias-free sample. However, all these methods presume that bias-free samples do exist in sufficient quantities and can be distinguished, which cannot always be guaranteed.

**Contrastive learning** is a self-supervised learning method proven to learn representations substantially beneficial to numerous downstream tasks, achieving state-of-the-art performance (Chen et al., 2020a;b; Grill et al., 2020; Chen & He, 2021; Khosla et al., 2020). Chen et al. (2020a) defines

the contrastive prediction task in two steps: a) augmenting two views from the same image by strong data augmentations (Bachman et al., 2019; Henaff, 2020; Krizhevsky et al., 2017) and b) maximizing agreement between the augmented views on the latent space by employing the contrastive loss (van den Oord et al., 2018) to capture core features of the image. In the process of maximizing the agreement between the augmented views, the encoder discards the deviating features between the views and learns transformation-invariant representations. For example, using color distortion methods such as *color jitter* and *gray scale* in the augmentation step encourages the encoder to learn color-invariant representations.

# 6. Conclusion

In this paper, we propose a novel debiasing method called Contrastive Debiasing via Generative bias-transformation (CDvG). This method is a general debiasing approach that does not require bias labels, bias type information, or even the existence of bias-free samples. Motivated by the observation that not only discriminative models but also image translation models tend to focus on bias, we utilize a translation model as a bias-translator to synthesize novel views with altered bias features. We then employ contrastive learning to pit one bias type against another and obtain bias-invariant representations. CDvG can be integrated with other debiasing techniques in a plug-and-play manner and especially shows good synergy when integrated with models that focus on bias-free samples, such as Nam et al. (2020). Our extensive experiment on various datasets shows that our method outperforms state-of-the-art, especially when bias-free samples are scarce or even absent.

As a future direction, and a current limitation, we aim to extend our method to datasets where bias-free data are the majority. Recent debiasing approaches, including our method, focus on the highly biased setting, which has yet to be conquered. While these situations may not require debiasing techniques from the start, suggesting a robust method that works in general is an important area of ongoing research.

# Acknowledgements

# References

Ahmed, F., Bengio, Y., Van Seijen, H., and Courville, A. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Bahng, H., Chun, S., Yun, S., Choo, J., and Oh, S. J. Learning de-biased representations with biased representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 528–539. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/bahng20a.html.

Brendel, W. and Bethge, M. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkfMWhAqYQ.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020a. URL http://proceedings.mlr.press/v119/chen20j.html.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Chen, X., Fan, H., Girshick, R. B., and He, K. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020b. URL https://arxiv.org/abs/2003.04297.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

.Gatys, L. A., Ecker, A. S., and Bethge, M. Texture and art with deep neural networks. *Current Opinion in Neurobiology*, 46:178–186, 2017. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2017.08.019. URL https://www.sciencedirect.com/science/article/pii/S095943881730065X. Computational Neuroscience.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*. OpenReview.net, 2019. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2019.html#GeirhosRMBWB19.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hong, Y. and Yang, E. Unbiased classification through bias-contrastive and bias-balanced learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=2OqZZAqxnn.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Kim, E., Lee, J., and Choo, J. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14992–15001, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Lee, J., Kim, E., Lee, J., Lee, J., and Choo, J. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021.

Li, Y., Li, Y., and Vasconcelos, N. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Loshchilov, I. and Hutter, F. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/mniha16.html.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Pezeshki, M., Kaba, O., Bengio, Y., Courville, A. C., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Sagawa*, S., Koh*, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Seo, S., Lee, J.-Y., and Han, B. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16742–16751, 2022.

Shrestha, R., Kafle, K., and Kanan, C. Occamnets: Mitigating dataset bias by favoring simpler hypotheses. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pp. 702–721. Springer, 2022.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pp. 1521–1528, USA, 2011. IEEE Computer Society. ISBN 9781457703942. doi: 10.1109/CVPR.2011.5995347. URL https://doi.org/10.1109/CVPR.2011.5995347.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, H., He, Z., and Xing, E. P. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJEjjoR9K7.

Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017a.

Zhu, Z., Xie, L., and Yuille, A. Object recognition with and without objects. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 3609–3615, 2017b. doi: 10.24963/ijcai.2017/505. URL https://doi.org/10.24963/ijcai.2017/505.

## A. Datasets

**Colored MNIST** is a biased version of MNIST with colors as biases.

**Biased MNIST** is a biased version of MNIST with multiple biases such as digit color, type of background texture, background color, co-occurring letters, and colors of the co-occurring letters.

**Corrupted CIFAR-10** is an artificially corrupted version of CIFAR-10 (Krizhevsky et al., 2009) to carry biases as proposed in Hendrycks & Dietterich (2019). Specifically, the dataset has been corrupted by the following types of method: {Snow, Frost, Fog, Brightness, Contrast, Spatter, Elastic transform, JPEG, Pixelate, and Saturate}.

**Biased Action Recognition (BAR)** is a real-world action dataset proposed by Nam et al. (2020). There are six action labels that are biased toward background places.

**BFFHQ** is proposed by Kim et al. (2021) which is curated from Flickr-Faces-HQ (Karras et al., 2019). It consists of face images where age (young/old) is a task label and gender (male/female) is a bias attribute.

**ImageNet-9 (IN-9)** is proposed by Bahng et al. (2020) which is a subset of ImageNet (Russakovsky et al., 2015) containing 9 super-classes with texture as biases (Ilyas et al., 2019).

**Waterbirds** is proposed by Sagawa* et al. (2020), which combines bird photographs from the Caltech-UCSD Birds-200-2011 (CUB) dataset with image backgrounds from the Places dataset. It consists of bird images where a type of bird is a task label and a background (water/land) is a bias attribute.

Colored MNIST, Corrupted CIFAR-10, and BFFHQ can be obtained from the official Github repository of DisEnt (Lee et al., 2021) (https://github.com/kakaoenterprise/Learning-Debiased-Disentangled).

Biased MNIST is available in the official Github repository of OccamNets (Shrestha et al., 2022) (https://github.com/erobic/occam-nets-v1).

BAR is available in the BAR Github repository (https://github.com/alinlab/BAR) provided by Nam et al. (2020).

ImageNet-9 is available in the official Github repository of ReBias (Bahng et al., 2020) (https://github.com/clovaai/rebias).

Waterbirds is available in the official Github respository of GroupDRO (Sagawa* et al., 2020) (https://github.com/kohpangwei/group_DRO).

## B. Implementation details

Basically, we follow the same experimental settings in baselines (Nam et al., 2020; Lee et al., 2021). All models are trained on 4 RTX-3090ti GPUs.

**Training StarGAN.** First, we specify the details of training biased StarGAN. We basically follow the default settings for architectures, optimizers weights for loss terms, and other training configurations in the Github repository of StarGAN (https://github.com/yunjey/stargan) across all datasets. For the generator, we use the basic architecture which is composed of total of 3 and 6 blocks for Colored MNIST and the others, respectively. Each block consists of 2 convolutional layers and a skip connection. For the discriminator, we use the architecture which is composed of total 4 and 5 blocks for {Colored MNIST, BFFHQ and IN-9} and {Corrupted CIFAR10 and waterbirds}, respectively. For Colored MNIST, we set the reconstruction weight to 500 and trained for 5000 iterations without random horizontal flipping. For IN-9 and Waterbirds, we resize them to 224 and use the original image size for others.

**Training Configuration.** For the encoder, we use MLP with three hidden layers for Colored MNIST, randomly initialized ResNet-18 (He et al., 2016) for Corrupted CIFAR-10 and BFFHQ We use batch sizes of 256, 128, and 64 for {Colored MNIST, and Corrupted CIFAR-10}, IN-9, and {BFFHQ and waterbirds}, respectively. We use learning rates of 0.001 for {Colored MNIST, Corrupted CIFAR-10, BFFHQ, IN-9 and waterbirds}.Also, We use the Adam optimizer with default parameters. We train the model for 120, 200, and 500 epochs for IN-9,{Colored MNIST, BFFHQ and waterbirds} and Corrupted CIFAR-10, respectively. We use cosine annealing from initial learning rates $lr$ to $lr * 0.1^3$ for learning rate scheduling (Loshchilov & Hutter, 2017) for all datasets. Note that we do not use random horizontal flipping for ColoredMNIST. Additionally, the original image size of BFFHQ is 128, but we resize them to 224 by following the previous work (Lee et al., 2021).

**Contrastive learning.** we use Normalized Temperature-scaled Cross Entropy *NT-Xent* (Chen et al., 2020a) with a temperature parameter 0.01. For projection head $H$, 2-layer MLP and a linear layer with the dimensions from the input to the output as [512, 512, 128] and [100, 100] for {Corrupted CIFAR-10, BFFHQ, IN-9 and waterbirds} and Colored MNIST respectively, following Chen et al. (2020a) for the most of the settings.

**Evaluation.** Note here that, by following Lee et al. (2021), the performances on BFFHQ whose task is a binary classification, are evaluated only on bias-free test samples that consist of young male and old female samples.

## C. Learning without Bias-free Samples

In Table 6, we empirically demonstrate that previous methods fail to effectively debias the model (showing low classification scores, almost comparable to Vanilla) when the bias-free samples are absent, i.e., when the biases are more malignant. Nam et al. (2020) and Lee et al. (2021) show surprisingly degraded performances due to the failure of reweighting scheme. Also, Kim et al. (2021) and Hong & Yang (2021) break down because they require bias-free samples necessarily to construct pairs of the bias-aligned and bias-free samples in the same class.

Table 6: Comparison of debiasing methods on three biased datasets devoid of bias-free samples (ratio 0%). We report averaged accuracy on the last epoch and the standard deviation over 3 runs.

| Method | ColoredMNIST | CorruptedCIFAR10 | BFFHQ |
|---|---|---|---|
| (ratio (%)) | 0 | 0 | 0 |
| Vanilla | 12.53±0.92 | 16.05±0.13 | 37.93±0.96 |
| LfF (Nam et al., 2020) | 13.16±1.87 | 15.88±0.45 | 39.67±1.00 |
| DisEnt (Lee et al., 2021) | 11.65±0.61 | 18.76±0.88 | 38.13±2.13 |
| ReBias (Bahng et al., 2020) | 12.72±0.07 | 14.40±0.49 | 38.73±0.38 |
| SoftCon (Hong & Yang, 2021) | 34.06±0.41 | 25.90±0.22 | 38.65±0.38 |
| **CDvG** | 95.97±0.18 | 31.24±0.31 | 42.80±0.33 |

Furthermore, to evaluate whether the reweighting scheme assigns sufficiently high weights to bias-free samples when bias-free samples are scarce, we further analyze DisEnt (Lee et al., 2021), which is the state-of-the-art, on Corrupted CIFAR10. In Figure 7, we measure the ratio of bias-free weight (blue bar), which implies how much the bias-free samples are focused, as the ratio of the sum of weights for bias-free samples to the total sum of weights for all samples $\frac{\sum_{i \in F} w_i}{\sum_{i=1}^{N} w_i}$ where $w_i$ is the assigned weight of $i$-th sample, $F$ is the set of indices of bias-free samples and $N$ is the total number of training samples. As the bias-free samples become scarce, the ratio of bias-free weight calculated by DisEnt decreases, which fails to debias the model, resulting in decreased accuracy (red line).

On the other hand, CDvG shows better performance even with 0% bias-free samples compared to the DisEnt with 0.5% bias-free samples. This is because CDvG does not necessarily require bias-free samples for debiasing, unlike the baselines (Bahng et al., 2020; Hong & Yang, 2021; Nam et al., 2020; Lee et al., 2021; Kim et al., 2021).
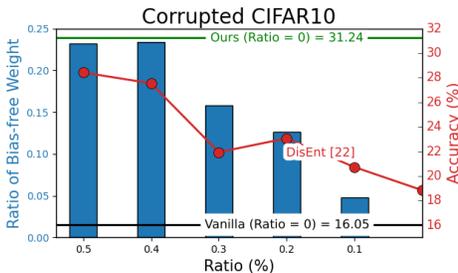


Figure 7: Changes in the ratio of bias-free weight (blue bar) and accuracy (red line) of DisEnt (Lee et al., 2021) as the ratio of bias-free samples (x-axis) decreases. The green and black horizontal lines are the accuracy of ours (GDvG) and Vanilla, respectively, when the bias-free samples are absent.

# D. Bias-transformed images by CycleGAN

As an example that other translation models also can be used in our framework, we show bias-transformed images generated by CycleGAN (Zhu et al., 2017a), which translates between two domains trained on BAR dataset.

The leftmost column of the figure contains the original images and each column is the transformed images of each target domain. The resulting images show the translated background bias attributes.
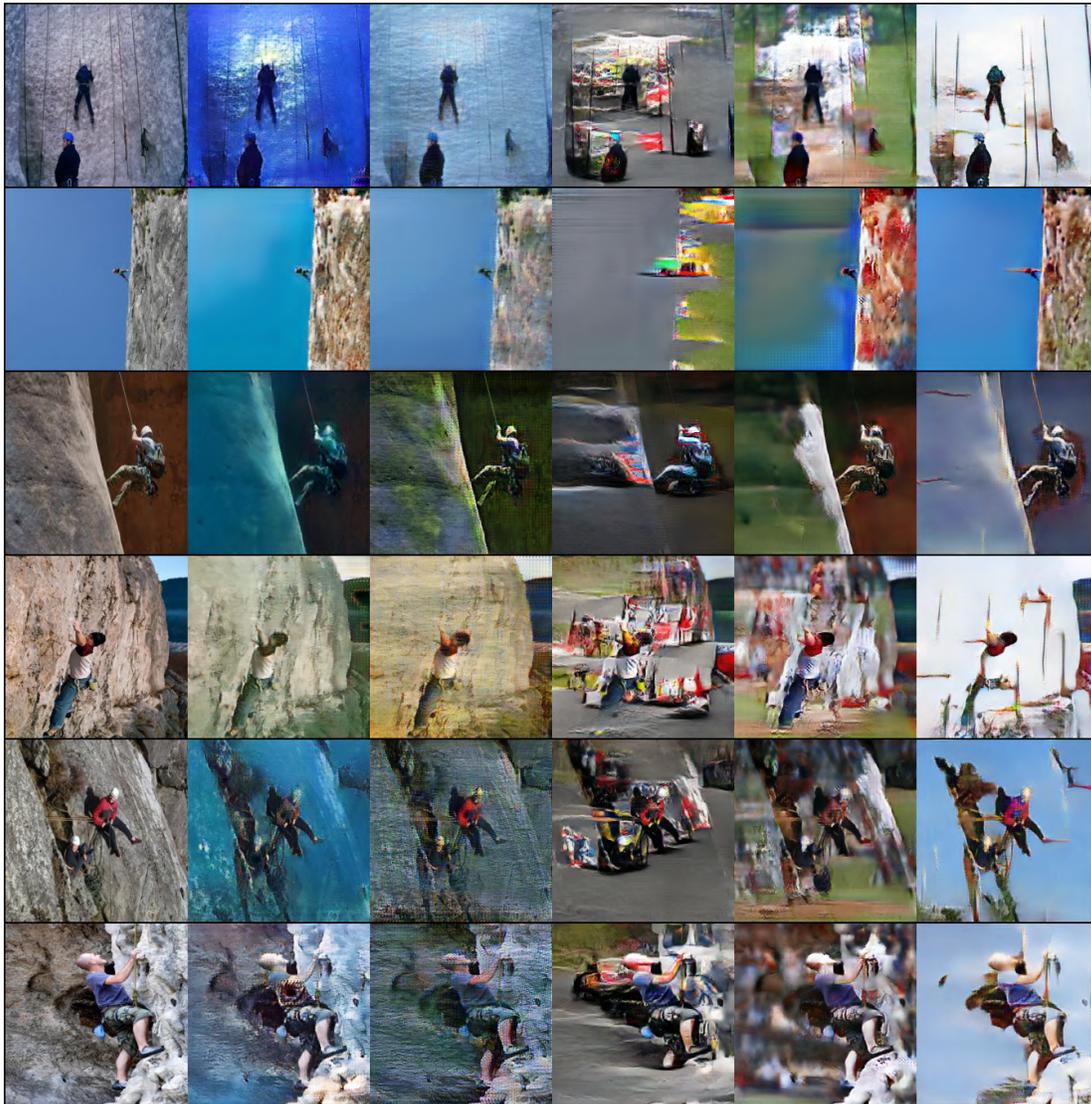


Figure 8: Bias-transformed images by CycleGAN on BAR dataset.

# E. StarGAN V2

We conducted an experiment in which the multi-task discriminator of Choi et al. (2020) was replaced with a domain discriminator. In Table 7, we empirically demonstrate that recent translation models are also inherently bias-susceptible to bias when using a domain classifier. In conclusion, the use of recent translation models can improve the debiasing performance by improving the quality of translated images, but there is a trade-off between cost and performance due to the large size of the network (6 times larger than StarGAN).

Table 7: Corrupted CIFAR-10 with Choi et al. (2020)

| Dataset | Ratio (%) | Vanilla | CDvG w/ StarGAN | CDvG w/ StarGANv2 |
|---------|-----------|---------|------------------|-------------------|
| Corrupted CIFAR-10 | 0.0 | 20.87 | 31.50 | **32.30** |
| | 5.0 | 41.12 | 42.75 | **47.80** |