A DATASETS

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). 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). Waterbirds is available in the official Github respository of GroupDRO (Sagawa et al., 2019).

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

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 to 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 an age (young/old) as a task label and a gender (male/female) as 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. (2019) which combines bird photographs from the Caltech-UCSD Birds-200-2011 (CUB) dataset with image backgrounds from the Places dataset. It consists of birds images where an type of bird as a task label and a background (water/land) as bias attribute.

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 3 and 6 blocks for Colored MNIST and the others, respectively. Each block consists of 2 convolutional layers and skip connection. For 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., 2015) for Corrupted CIFAR-10 and BFFHQ We use batch sizes of 256, 128, 64 for {Colored MNIST, and Corrupted CIFAR-10}, IN-9 and {BFFHQand 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 5, 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) shows 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 5: 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 analyse DisEnt (Lee et al., 2021), which is the state-of-the-art, on Corrupted CIFAR10. In Figure 5, 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).

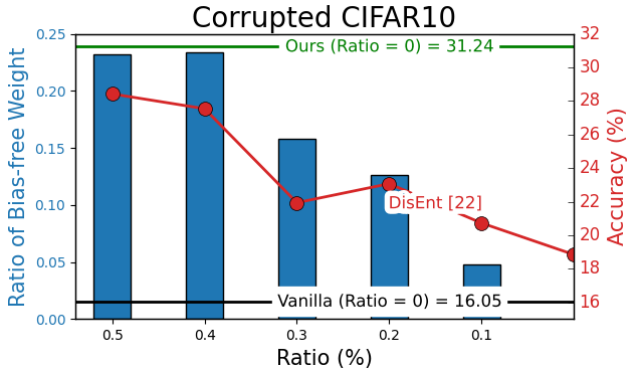

Figure 5: Changes in the ratio of bias-free weight (blue bar) and accuracy (red line) of DisEnt (Lee et al., 2021) as the ratio of bias-free samples (x-axis) decreases. The green and black horizontal lines are the accuracy of ours (GDvG) and Vanilla, respectively, when the bias-free samples are absent.

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.

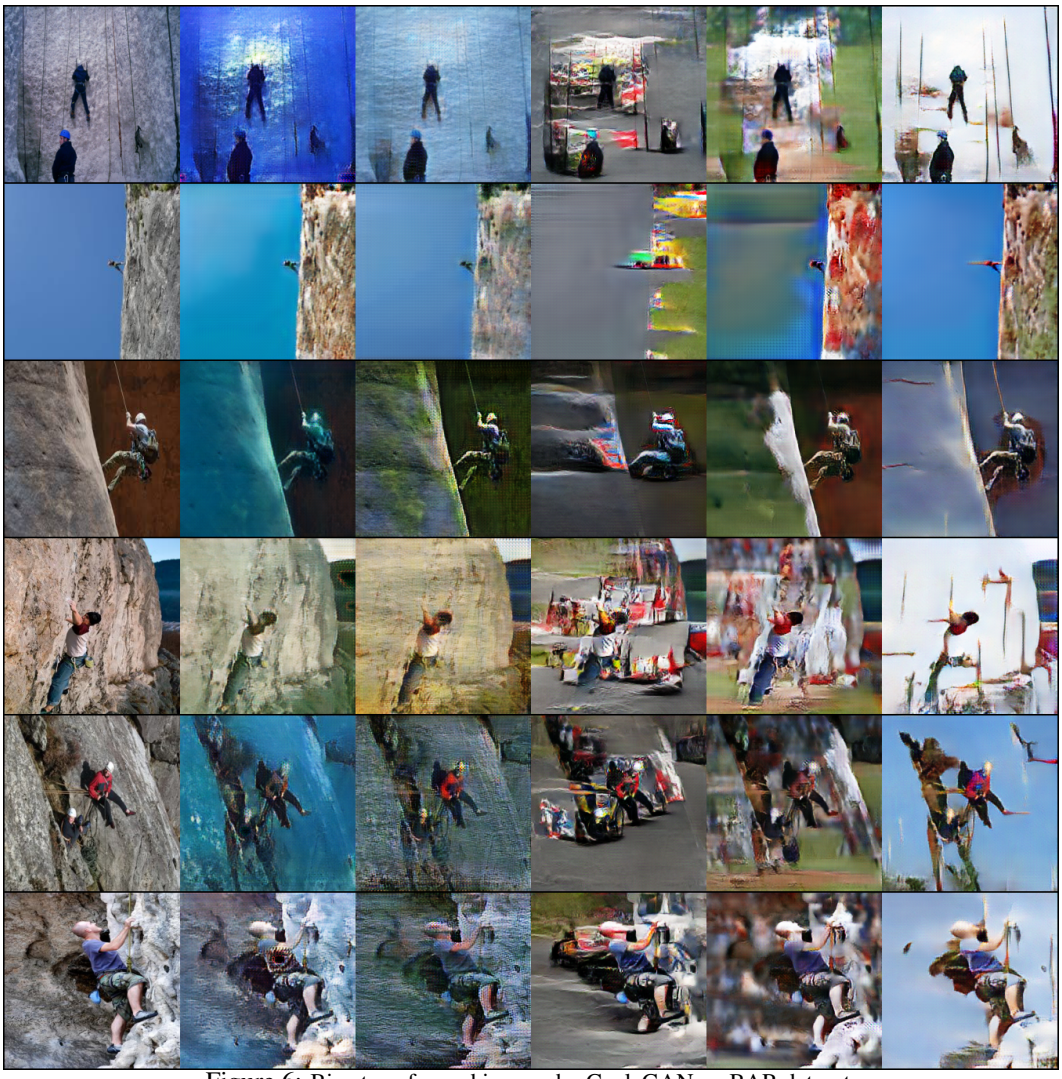

Figure 6: Bias-transformed images by CycleGAN on BAR dataset.