Supplementary Materials: Hybrid Sample Synthesis-based Debiasing of Classifier in Limited Data Setting

Piyush Arora*
Indian Institute of Technology Jodhpur, India
arora.8@iitj.ac.in

Pratik Mazumder*
Indian Institute of Technology Jodhpur, India
pratikm@iitj.ac.in

1. Datasets and Additional Implementation Details

We utilized the following datasets in our experiments: Colored MNIST, Corrupted CIFAR-10, and BFFHQ.

Colored MNIST [6] is a variant of the original MNIST dataset [1] with color bias. A specific color is injected into the images of the MNIST dataset with some perturbation. Each digit in Colored MNIST is associated with a specific foreground color. Consequently, the naive baseline models learn to classify some digits based on only the color of the digit, which is an unwanted correlation/bias that such models learn.

Corrupted CIFAR-10, on the other hand, consists of ten different types of texture biases applied to the CIFAR-10 dataset [4]. This dataset was constructed following the design protocol of Hendrycks and Dietterich [2], and each class is highly correlated with a particular texture. Specifically, Corrupted CIFAR-10 Type 0 contains texture biases such as snow, frost, fog, brightness, contrast, spatter, elastic, JPEG, pixelate, and saturate, while Corrupted CIFAR-10 Type 1 contains texture biases such as Gaussian noise, shot noise, impulse noise, speckle noise, Gaussian blur, defocus blur, glass blur, motion blur, zoom blur, and original.

The Biased FFHQ (BFFHQ) dataset was derived from the FFHQ dataset [3] by the authors of [5], which consists of human face images labeled with various facial attributes. The BFFHQ comprises age and gender as the intrinsic and biased attributes, respectively, and compiles a collection of images that exhibit a strong correlation between these two attributes. In order to achieve this bias, a majority of the images of males have males with ages between 40 and 59, and a majority of the images of females have females with ages between 10 and 29. Therefore, a majority of the females in the dataset are young, while a majority of the males in the dataset are old. The bias-aligned samples of the dataset are the samples that follow this bias.

In this work, we use the PyTorch framework [7] and Python 3.0 for all the experiments. We use the NVIDIA RTX A5000 graphics processing unit for our experiments. We run all the experiments 3 times with different seeds and report the average accuracy and the 95% confidence score.

2. Additional Experiments with p = 1% and 2%

p	Vanilla	LfF	LDD	DebiAN	Ours
2%	57.04	65.58	63.18	56.54	69.08
1%	47.51	52.33	51.02	47.43	55.72

Table 1. Additional Experimental Results on reduced CMNIST for $\sigma = 0.05$

We perform additional experiments with a significantly limited amount of training data, i.e., p=1% and 2% of the training data. The results for these experiments on the reduced Colored MNIST dataset with $\sigma = 0.05$ are reported in Table 1. The results indicate that the performance of the vanilla model falls significantly as compared to the results in the main paper as the amount of training data decreases further. The results indicate that the effectiveness of the bias mitigation approaches decreases even further. The results also indicate that the proposed approach significantly outperforms the vanilla, LfF, LDD, and DebiAN approaches by absolute margins of 12.04%, 3.5%, 5.9%, 12.54%, respectively, for p = 2% and by absolute margins of 8.21%, 3.39%, 4.7%, 8.29%, respectively, for p=1%. Therefore, the proposed approach is more effective at debiasing models as compared to the other approaches in this setting.

References

[1] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 1

[2] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019. 1

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[5] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 25123–25133, 2021. 1

[6] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020. 1

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch, 2017. 1