Supplementary Material *for* Controllable Feature Whitening for Hyperparameter-Free Bias Mitigation

A. Datasets

Corrupted CIFAR-10 was proposed by [9], and constructed by corrupting the CIFAR-10 dataset [5]. Consequently, this dataset is annotated with category of object and type of corruption used. Each object class is highly correlated with a certain type of corruption. We select the object type and corruption type as the target and bias attributes, respectively. We conduct experiments by varying the ratio of *bias-conflicting* samples by selecting from {0.5%, 1.0%, 2.0%, 5.0%}.

Biased FFHQ was proposed by [2] and curated from FFHQ datset [1]. It contains human face images that annotated with the gender and age. In the bFFHQ, 99.5% of the women are young (age: 10-29), and 99.5% of the men are old (age: 40-59). Therefore, the ratio of the *bias-conflicting* samples of the bFFHQ is 0.5%. We select the age as the target attribute, and the gender as the bias attribute.

Celeb-A is a large-scale face attributes dataset [8]. It contains a total of 202,599 images annotated with 40 binary attributes and 5 landmark location. Following the official train-test split, we train with 162,770 training images and evaluate accuracy on 19,962 test images. Following [9] [3], we select *BlondHair*, *HeavyMakeup*, *Attractiveness*, *Bignose*, *Bag-under-eyes*, *Male*, and *Young* as the attribute candidates, and choose the highly correlated target and bias attributes among them. For example, in Celeb-A dataset, the most of the male images doesn't have a blond hair or a heavy makeup.

WaterBirds [11] is a dataset in which the target attribute is bird species (waterbird vs. landbird) and the bias is the background (water or land). WaterBirds dataset is constructed by combining bird with backgrounds in a biased way: most waterbirds are pictured on water (*bias-aligned*) and most landbirds on land, while only a 5% of images are placed in the opposite background (*bias-conflicting*).

B. Implementation Details

Training Configuration. We follow the training settings of the previous works, LfF [9], DisEnt [6] and CSAD [15]. We employ Pytorch `torchvision` implementations of the ResNet-18 and ResNet-50 as the encoder network. We train the network with Adam optimizer with the default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and weight decay of 0. We set the decaying step to 10K. For Celeb-A and WaterBirds, following previous works [9, 12, 15], we employ ImageNet pretrained ResNet-18 and ResNet-50 which are provided by Pytorch `torchvision`. Experiments were run on NVIDIA Titan Xp GPUs.

C. Evaluation Metrics

Let $(\hat{Y}, Y, B) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{B}$ denote the prediction, target attribute, and bias attribute, respectively. The *worst-group* test accuracy can be expressed using the following equation:

$$acc_{wg} = \min_{y,b \in \mathcal{Y} \times \mathcal{B}} p(\hat{Y} = y | Y = y, B = b). \tag{1}$$

Then, we can express the *unbiased* and *bias-conflicting* test accuracy using the following equation:

$$acc = \frac{1}{|\Omega|} \sum_{(y,b) \in \Omega} p(\hat{Y} = y | Y = y, B = b), \tag{2}$$

where Ω is the set of the target-bias pairs. To calculate the *unbiased* test accuracy, we average the test accuracy over the all possible target-bias pairs (i.e., $\Omega = \mathcal{Y} \times \mathcal{B}$). To calculate the *bias-conflicting* test accuracy, we average the test accuracy over the *bias-conflicting* target-bias pairs (e.g., old-woman, young-man in bFFHQ, and *BlondHair* Male, not *BlondHair* Female in Celeb-A).

Method	Backbone	Bias Label	*Worst-G*	*Mean*
Vanilla	Res50	✗	74.9±2.4	**98.1**±**0.1**
LfF [9]	Res50	✗	78	91.2
JTT [7]	Res50	✗	86.7	93.3
SSA [10]	Res50	△	89.0±0.6	92.2±0.9
CNC [14]	Res50	△	88.5±0.3	90.9±0.1
$\text{DFR}^{\text{Val}}_{\text{Tr}}$ [4]	Res50	△	92.9±0.2	94.2±0.4
GroupDro [11]	Res50	✓	91.4	93.5
LISA [13]	Res50	✓	89.2	91.8
Ours	Res50	✓	**93.46**±0.11	96.82±0.52
Vanilla	Res50(w/o IN)	✗	6.9±3.0	**88.0**±**1.1**
$\text{DFR}^{\text{Val}}_{\text{Tr}}$ [4]	Res50(w/o IN)	△	56.70±1.3	61.03±1.62
Ours	Res50(w/o IN)	✓	**63.56**±**0.76**	69.72±1.14

Table 1. Comparison of the *Mean* and *Worst-Group* test accuracy (%) on the WaterBirds. Best performing results are marked in bold, while the second-best results are denoted with underlines. Res50(w/o IN) referes to ResNet-50 without ImageNet pretraining.

D. Results on WaterBirds

In Table 1, we report the *Mean* and *Worst-Group* test accuracy on WaterBirds [11]. By following [4, 11], we compute *Mean* accuracy by weighting the group accuracies according to their prevalence in the training data. For *Worst-Group* accuracy, our method consistently outperforms all competing algorithms, demonstrating its effectiveness in mitigating bias. For *Mean* accuracy, our method achieves the best performance after *Vanilla*, which performs well on the *bias-aligned* samples but fails on *bias-conflicting* samples. These results confirm that our approach achieves both superior fairness and overall performance. Notably, the performance gap between our method and DFR, the second-best performing approach, increases when the backbone network is not pretrained on ImageNet. This suggests that despite the need for bias labels, our method remains more robust and broadly applicable across different settings.

References

[1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[2] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14992–15001, 2021. 1

[3] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. *Advances in Neural Information Processing Systems*, 35:18403–18415, 2022. 1

[4] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 2

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 1

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

[7] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 2

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 1

[9] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 1, 2

[10] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022. 2

[11] 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. 1, 2

[12] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13508–13517, 2021. 1

[13] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022. 2

[14] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 2

[15] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15002–15012, 2021. 1