Supplementary Material for Bias Mimicking: A Simple Sampling Approach for Bias Mitigation

Maan Qraitem[1], Kate Saenko[1,2], Bryan A. Plummer[1]
[1]Boston University [2]MIT-IBM Watson AI Lab
{mqraitem,saenko,bplum}@bu.edu

		BM	BM+US	BM+UW	BM+OS
UTK-Face	UA	79.7±0.4	79.2±1.0	79.9±0.2	79.3±0.2
Age	BC	79.1±2.3	77.6±0.9	77.5±1.7	78.7±1.6
UTK-Face	UA	90.8±0.2	90.9±0.5	91.1±0.2	90.7±0.4
Race	BC	90.7±0.5	91.1±0.3	91.6±0.1	90.9±0.5
CelebA	UA	90.8±0.4	91.1±0.2	91.1±0.4	91.1±0.1
Blonde	BC	87.1±0.6	87.9±0.3	87.9±0.7	87.7±0.4
CIFAR-S	UA	91.6±0.1	91.7±0.1	91.8±0.0	91.6±0.2
	BC	91.1±0.1	91.2±0.2	91.4±0.2	91.2±0.2

Table 1. **Sampling for multi-class prediction head** compare the effects of using different sampling methods to train the multi-class prediction in our proposed method: Bias Mimicking. We underline results where sampling methods make significant improvements. Refer to Section A for discussion.

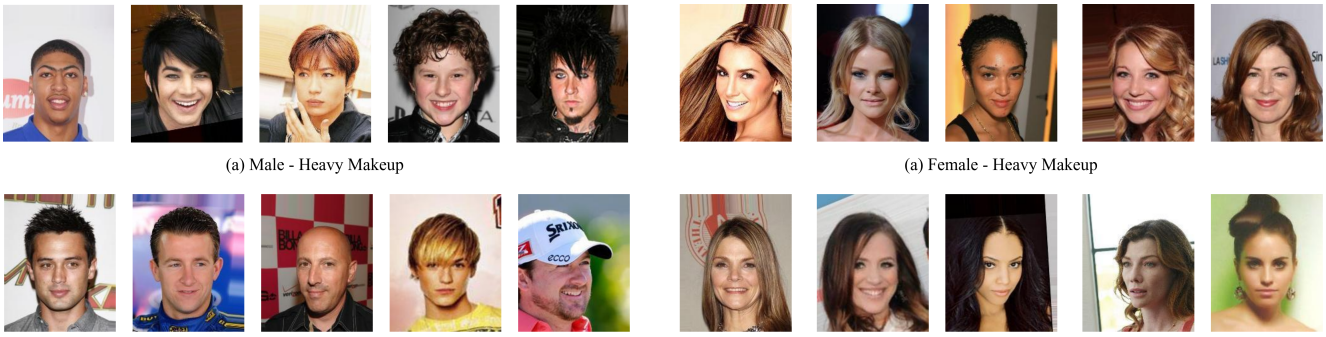
A. Sampling methods Impact on Multi-Class Classification Head

Bias Mimicking produces a binary version d_c of the dataset D for each class c. Each d_c preserves class c samples while undersampling each c' such that the bias within c' mimics that of c. A debiased feature representation is then learned by training a binary classifier for each d_c. When the training is done, using the scores from each binary predictor for inference is challenging. This is because each predictor is trained on a different distribution of the data, so the predictors are uncalibrated with respect to each other. Therefore, to perform inference, we train a multi-class prediction head using the learned feature representations and the original dataset distribution. Moreover, we prevent the gradients from flowing into the feature space since the original distribution is biased. Note that we rely on the assumption that the correlation between the target labels and bias labels are minimized in the feature space, and thus the linear layer is unlikely to relearn the bias. During our experiments outlined in Section 4, we note that this approach was sufficient to obtain competitive results. This section explores whether we can improve performance by using sampling methods to train the linear layer. To that end, observe results in Table 1. We underline the rows where the sampling methods make improvements. We note that the sampling methods did not improve performance for three of the four benchmarks in our experiments. However, on CelebA, we note that the sampling methods marginally improved performance. We suspect this is because a small amount of the bias might be relearned when training the multi-class prediction head since the input distribution remains biased.

B. Heavy Makeup Benchmark

Prior work [3] uses the Heavy Makeup binary attribute prediction task from CelebA [6] as a benchmark for bias mitigation, where Gender is the sensitive attribute. In this experiment, Heavy Makeup's attribute is biased toward the sensitive group: Female. We note that the notion of "Heavy Makeup" is quite subjective. The attribute labels may vary significantly according to cultural elements, lighting conditions, and camera pose considerations. Thus, we expect a fair amount of label noise, *i.e.*, inconsistency with label assignment. We document this problem in a Quantitative and Qualitative analysis below.

Quantitative Analysis: We randomly select a total of 200 pairs of positive and negative images. We ensure the samples are balanced among the four possible pairings, *i.e.*, (Heavy Makeup-Male, Non Heavy Makeup-Female), (Heavy Makeup-Female, Non Heavy Makeup-Male), (Heavy Makeup-Male, Non Heavy Makeup-Male), (Heavy Makeup-Female, Non Heavy Makeup-Female). We asked three independent annotators to label which image in the pair is wearing "Heavy Makeup". Then, we calculate the percentage of disagreement between the three annotators and the ground truth labels in the dataset. We note that $32.3\% \pm 0.02$ of the time, the annotators on average disagreed with the ground truth. The noise is further amplified when the test set used in [3] is examined. In particular, Male-Heavy Make up (an under-represented subgroup)



(a) Male - Heavy Makeup (a) Female - Heavy Makeup

(b) Male - Non Heavy Makeup (b) Female - Non Heavy Makeup

Figure 1. Randomly sampled images from the four subgroups: Female-Heavy Makeup, Female-Non-Heavy Makeup, Male-Heavy Makeup, and Male-Non-Heavy Makeup in CelebA dataset [6]. Note the that there is not a clearly differentiating signal for the attribute Heavy Makeup. Refer to Section B for discussion.

		Nonsampling Methods					Sampling Methods				
		Vanilla	Adv [2]	G-DRO [7]	DI [9]	BC+BB [3]	OS [9]	UW [1]	US [4]	BM	BM + OS
CelebA	UA	90.9±0.1	90.7±0.2	93.2±0.1	92.0±0.1	92.4±0.1	92.4±0.3	92.8±0.0	92.6±0.1	92.7±0.1	92.7±0.1
Smiling	BC	84.3±0.2	84.7±0.4	92.2±0.1	91.3±0.2	92.6±0.1	91.5±0.2	92.4±0.2	92.1±0.2	92.3±0.2	92.2±0.1
CelebA	UA	86.3±0.7	87.1±0.3	88.5±0.2	86.7±0.7	87.7±0.1	87.6±0.3	88.5±0.1	88.4±0.2	87.6±0.7	88.5±0.1
Black Hair	BC	82.7±0.6	83.4±0.5	88.3±0.4	86.6±1.2	86.6±0.3	85.6±0.6	88.0±0.2	87.3±0.1	87.8±1.3	88.5±0.7
Average	UA	88.6±0.4	88.9±0.2	**90.8±0.1**	89.3±0.4	90.0±0.1	90.0±0.3	**90.6±0.1**	**90.5±0.1**	**90.2±0.4**	**90.6±0.1**
	BC	83.5±0.4	84.0±0.4	**90.2±0.2**	88.9±0.7	89.6±0.2	88.5±0.4	**90.2±0.2**	89.7±0.1	**90.0±0.7**	**90.3±0.4**

Table 2. Expanded benchmarks from the CelebA dataset [6]. Refer to Section C for discussion.

only contains 9 testing samples. We could not visually determine whether 4 out of these 9 images fall under Heavy Makeup. Out of the 5 remaining images, 3 are of the same person from different angles. Thus, given the noise in the training set, the small size of the under-represented group in the test set, and its label noise, we conclude that results from this benchmark will not be reliable and exclude it from our experiments.

Qualitative Analysis: We sample random 5 images from the following subgroups: Female-Heavy Makeup, Female-Non-Heavy Makeup, Male-Heavy Makeup, and Male-Non-Heavy Makeup (Fig 1). It is clear from the Figure that there is no firm agreement about the definition of Heavy Makeup.

C. Additional Benchmarks

In Section B, we note that the CelebA attribute Heavy-Makeup usually used in assessing model bias in prior work [3, 8] is a noisy attribute, *i.e.*, labels are inconsistent. Therefore, we choose not to use it in our experiments. Alternatively, we provide results on additional attributes where labels are more likely to be consistent. To that end, we choose to classify the attributes: Smiling and Black Hair, where Gender is the bias variable. The original distribution of each attribute is not sufficiently biased with respect to Gender to note any significant change in performance. Thus, we subsample each distribution to ensure that each attribute is biased toward Gender. We provide the splits for the resulting distributions in the attached code base. Refer for Table 2 for results.

Note that our method Bias Mimicking performance marginally lags behind other methods when predicting "Black Hair" attribute. However, when the multi-class prediction layer is trained with an oversampled distribution (BM + OS), then the gap is bridged. This is consistent with the observation in Table 1 where oversampling marginally improves our method performance on CelebA. These observations indicate that on some benchmarks, a small amount of the bias might be relearned through the multi-class prediction head. To ensure that this bias is mitigated, it is sufficient to oversample the input distribution. Moreover, since oversampling the input distribution does not change performance on other datasets as indicated in Table 1, we recommend that the input distribution for the multi-class prediction head is oversampled to ensure the best performance.

Overall, (BM + OS) performs comparably to sampling

	Learning Rate	Weight Reg	Group Adjustment
UTK-Face Age	0.001	0.01	4
UTK-Face Race	0.001	0.001	4
CelebA Blonde	0.001	0.1	3
CelebA Smiling	0.0001	0.01	2
CelebA Black Hair	0.0001	0.01	3
CIFAR-S	0.01	0.01	5

Table 3. Hyperparameters used for GroupDRO [7]. Refer to Section D for further discussion.

	US [4]	OS [9]	Other Methods
UTK-Face Age	400	7	20
UTK-Face Race	120	10	20
CelebA Blonde	170	4	10
CelebA Black Hair	40	5	10
CelebA Smiling	30	5	10
CIFAR-S	2000	100	200

Table 4. Number of Epochs used to train each method. Refer to Section D for further discussion.

and nonsampling methods. This is consistent with our results on CelebA dataset in Section 4 of the main paper where we predict "Blonde Hair". More concretely, Undersampling performs comparably and sometimes better than nonsampling methods. This is reaffirming that predicting attributes on CelebA is relatively an easy task that dropping samples to balance subgroup distribution is sufficient to mitigate bias. However, as discussed in Section 4 of the main paper, vanilla sampling methods (Undersampling, Upweighting, Oversampling) perform poorly on some datasets. For example, as we note in Table 1 in the main paper, Undersampling performs considerably worse than nonsampling methods on the Utk-Face dataset as well the CIFAR-S dataset. Moreover, Upweighting performs substantially worse on CIFAR-S. Finally, Oversampling performs consistently worse on every benchmark. However, only our method, Bias Mimicking, manages to maintain competitive performance with respect to nonsampling methods on all datasets.

D. Model and Hyper-parameters Details

We test bias Mimicking on six benchmarks. Three Binary Classification tasks on CelebA [6], namely, Blonde, Black Hair, and Smiling, Two Binary Classification tasks on UTK-Face [10], namely Race and Age and one multiclass task CIFAR-S. We provide further info below.

	100%	75%	50%	25%
UTK-Face Age	**79.7**	78.9	78.0	76.7
UTK-Face Race	**90.8**	90.6	89.9	88.3
CelebA Blonde	**90.8**	90.3	90.1	89.9

Table 5. Comparing the performance of our method's Unbiased Accuracy (UA) where we use $x\%$ of the linear program solution. Refer to E for discussion.

Optimization Following [3], we use ADAM [5] optimizer with learning rate 0.0001 on CelebA and UTK-Face. For CIFAR-S, following [9], we use SGD with learning rate 0.1. GroupDRO [7], however, has not been tuned before on the benchmarks in our study. Even for CelebA Blonde, the method was not tuned on the more challenging split in this study. Therefore, we grid search the learning rate/weight regularization/group adjustment and choose the best over the validation set. Refer to Table 3 for our final choices. With respect to BC+BB, the method was not benchmarked on CIFAR-S. Therefore, we run a grid search over the method's hyperparameters and choose $\alpha = 1.0$, $\gamma = 10$. Finally, as discussed in Section 4, UW struggles to optimize over CIFAR-S with learning rate 0.1. Therefore, we tune the learning rate and we find that 0.0001 to work the best over the validation set.

Total Number of Epochs As noted in Section 4.1 in the paper, a model trained with Undersampling sees fewer iterations than baselines per epoch and a model trained with Oversampling sees more iterations per epoch. Therefore, we adjust the number of epochs for both methods such that the total number of iterations seen by the model is the same across all methods tested in our experiments. Refer to Table 4 for a breakdown of the total number of epochs used to train each method.

Augmentations: For all benchmarks, we augment the input images with a horizental flip. BC+BB [3] uses extra augmentation functions. Refer to [3] for further details.

Splits: Note on CelebA, unlike [3], we use CelebA validation set for validation and test set for testing rather than using the validation set for testing and a split of the training set for validation.

E. Effect of the Linear Program Constraint

As discussed in Section 3 in the paper, We use a linear program to determine how the training distribution is subsampled to mimick the bias. The program is constrained such that the resulting distributions preserve the most number of samples because fewer retained samples may compromise the model performance. To verify this the importance of this step, we train our model using distributions that maintain x% of the Linear Program solution where $x < 100$. Note the results in Table 5. Note how the performance drops emphasizing the importance of this constraint.

References

[1] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 872–881. PMLR, 09–15 Jun 2019. 2

[2] Judy Hoffman, Eric Tzeng, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4068–4076, 2015. 2

[3] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2, 3

[4] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, pages 429–449, 2002. 2, 3

[5] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 3

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 2, 3

[7] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020. 2, 3

[8] 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 (CVPR)*, pages 13508–13517, June 2021. 2

[9] Zeyu Wang, Klint Qinami, Ioannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[10] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017. 3