Appendix for “Poisoning for Debiasing: Fair Recognition via Eliminating Bias Uncovered in Data Poisoning”

Table 2: The Model Bias (in %, ↓), Avg. Group Accuracy (in %, ↑), and Worst Group Accuracy (in %, ↑) of models trained on CelebA. Here *bn*, *a*, *bl* and *bu* respectively denote *bignose*, *attractive*, *blonde*, *bags-under-eyes*. The best results with unknown biases are highlighted in bold. * indicates that the method knows the bias label of training samples.

Method	T=*bn*			T=*a*			T=*bl*			T=*bu*		
	ModelBias ↓	AvgACC ↑	WorstACC ↑	ModelBias ↓	AvgACC ↑	WorstACC ↑	ModelBias ↓	AvgACC ↑	WorstACC ↑	ModelBias ↓	AvgACC ↑	WorstACC ↑
Vanilla	31.40±0.6	71.01±0.3	43.36±0.7	23.51±0.3	74.13±0.2	62.34±0.4	32.05±0.3	82.83±0.2	58.32±0.5	17.70±0.4	71.51±0.3	43.71±0.6
Focal	23.29±0.5	70.99±0.2	47.61±0.4	24.25±0.4	76.56±0.2	61.36±0.5	30.05±0.3	83.78±0.2	64.35±0.4	16.31±0.3	70.87±0.3	42.47±0.6
LfF	16.73±0.4	68.42±0.5	53.12±0.6	17.53±0.7	76.57±0.4	67.36±0.9	29.75±0.6	75.32±0.6	49.10±1.1	18.73±0.4	70.53±0.5	43.51±0.8
JTT	14.29±0.5	72.31±0.3	55.09±0.7	15.06±0.5	77.61±0.4	65.33±0.8	13.07±0.5	85.02±0.2	75.53±0.5	15.34±0.4	70.22±0.2	54.01±0.3
DebiAN	29.03±0.6	69.41±0.3	39.75±0.8	22.39±0.5	76.63±0.3	59.22±0.9	29.35±0.3	77.29±0.3	63.81±0.7	19.38±0.4	70.65±0.3	44.95±0.5
Echoes	19.95±0.6	66.19±0.4	42.73±0.9	27.52±0.4	72.57±0.3	65.11±0.8	18.27±0.5	76.53±0.2	63.52±0.7	16.52±0.7	70.52±0.4	51.23±0.8
BE	15.57±0.9	69.57±0.5	56.54±1.3	16.21±0.6	76.74±0.4	67.14±0.7	15.84±0.5	80.58±0.4	69.21±0.7	14.32±0.3	71.65±0.6	59.14±0.8
Poisoner	**6.56**±0.5	**74.49**±0.2	**69.61**±0.5	**3.57**±0.3	**79.98**±0.2	**77.04**±0.4	**8.05**±0.4	**91.37**±0.3	**85.78**±0.4	**7.61**±0.3	**76.58**±0.2	**68.29**±0.3
GroupDRO*	5.54±0.3	74.97±0.2	66.28±0.4	4.02±0.2	79.72±0.2	76.83±0.3	7.61±0.2	92.51±0.1	81.97±0.2	7.96±0.3	77.81±0.1	67.00±0.4

Table 1: Performance of Poisoner* on the CelebA (T=*bn*) dataset.

Method	Avg Acc (↑)	Worst Acc (↑)	Bias (↓)	Bias Accuracy (↑)
Vanilla	71.01±0.3	43.36±0.7	31.40±0.6	-
Poisoner	74.49±0.2	69.61±0.5	6.56±0.5	92.26±0.5
*Poisoner**	74.61±0.1	70.53±0.4	6.39±0.6	93.34±0.6

A CORRECTIONS TO THE MAIN PAPER

We apologize for the use of two different forms to represent the same concept in the main paper. Specifically, the symbol D_{con} in Lines 378 and 387 should be corrected to d_{con}. We apologize for any reading difficulties this may have caused.

B ANOTHER METHOD FOR IDENTITY BIAS-CONFLICTING SAMPLES

In the main paper:

> For each sample x_i in the current batch, we compare x_i with the other data in the mini-batch $\mathcal{B}$ as follows:
>
> $$d_{\text{con}}\left(x_i;\theta\right) = -\frac{1}{|J(i)|}\sum_{j\in J(i)} log\frac{\exp\left(z_i\cdot z_j\right)}{\sum_{a\in A(i)}\exp\left(z_i\cdot z_a\right)} \quad (1)$$
>
> The larger the $d_{\text{con}}\left(x_i;\theta\right)$, the more likely it is that the sample x_i is a bias-conflicting sample of the current target class.

The above approach identifies bias-conflicting samples in the model’s training on mini-batch $\mathcal{B}$, offering a resource-efficient advantage.

However, this method requires at least two samples with the same label in $\mathcal{B}$, which might not be feasible in cases with a large number of labels (e.g., 1000 classes). To address this situation, we propose using comparison data from outside the mini-batch $\mathcal{B}$.

Specifically, we sampled a larger batch $\mathcal{B}'$ to support comparisons with samples in $\mathcal{B}$. Thus, the calculation of $d_{\text{con}}\left(x_i;\theta\right)$ for samples in $\mathcal{B}$ is based on both $\mathcal{B}$ and $\mathcal{B}'$.

We implemented our new approach, called Poisoner*, on CelebA (T=*bn*) and present comparative results in Table 1, demonstrating that Poisoner* offers debiasing performance consistent with Poisoner.

Table 3: The debiasing performance on four benchmark datasets with general bias. * indicates that the method knows the bias label of training samples. - denotes that the test set is not applicable for evaluating model bias.

Method	WaterBirds			Dogs & Cats			C-MNIST^1			C-MNIST^2			ImageNet-B		
	ModelBias ↓	AvgACC ↑	WorstACC ↑	ModelBias ↓	AvgACC ↑	WorstACC ↑	ModelBias ↓	AvgACC ↑	WorstACC ↑	ModelBias ↓	AvgACC ↑	WorstACC ↑	ModelBias ↓	AvgACC ↑	WorstACC ↑
Vanilla	35.71$_{\pm0.5}$	78.17$_{\pm0.4}$	46.61$_{\pm0.7}$	-	50.35$_{\pm1.3}$	47.76$_{\pm1.2}$	43.27$_{\pm2.7}$	56.28$_{\pm2.8}$	10.30$_{\pm3.3}$	32.73$_{\pm2.8}$	67.09$_{\pm1.3}$	7.02$_{\pm1.7}$	67.09$_{\pm0.7}$	61.37$_{\pm0.3}$	23.60$_{\pm0.9}$
Focal	30.12$_{\pm0.5}$	77.99$_{\pm0.3}$	56.71$_{\pm0.6}$	-	68.25$_{\pm1.0}$	65.85$_{\pm1.2}$	43.82$_{\pm1.7}$	56.91$_{\pm1.3}$	3.05$_{\pm1.2}$	29.61$_{\pm0.2}$	70.10$_{\pm0.2}$	9.01$_{\pm0.4}$	66.99$_{\pm0.7}$	61.16$_{\pm0.4}$	19.30$_{\pm0.8}$
LfF	13.57$_{\pm0.5}$	79.64$_{\pm0.3}$	63.56$_{\pm0.6}$	-	72.91$_{\pm0.5}$	50.10$_{\pm0.6}$	12.99$_{\pm0.8}$	75.26$_{\pm0.7}$	28.41$_{\pm0.9}$	11.29$_{\pm1.0}$	83.92$_{\pm0.5}$	59.57$_{\pm1.2}$	37.97$_{\pm0.7}$	64.21$_{\pm0.5}$	28.20$_{\pm0.8}$
JTT	12.06$_{\pm0.8}$	75.61$_{\pm0.4}$	58.19$_{\pm0.9}$	-	73.95$_{\pm0.5}$	67.45$_{\pm0.6}$	13.07$_{\pm0.6}$	74.05$_{\pm0.5}$	30.03$_{\pm0.8}$	13.38$_{\pm0.4}$	78.34$_{\pm0.3}$	56.10$_{\pm0.5}$	35.24$_{\pm0.7}$	65.38$_{\pm0.6}$	30.54$_{\pm0.8}$
DebiAN	12.36$_{\pm0.6}$	77.72$_{\pm0.5}$	59.22$_{\pm0.7}$	-	71.24$_{\pm0.5}$	67.68$_{\pm0.7}$	17.08$_{\pm0.4}$	70.48$_{\pm0.4}$	36.18$_{\pm0.5}$	15.25$_{\pm0.6}$	79.51$_{\pm0.5}$	57.61$_{\pm0.8}$	44.58$_{\pm0.6}$	62.40$_{\pm0.5}$	29.32$_{\pm0.6}$
Echoes	14.52$_{\pm0.5}$	78.79$_{\pm0.2}$	62.73$_{\pm0.5}$	-	84.56$_{\pm0.4}$	82.17$_{\pm0.8}$	16.27$_{\pm0.4}$	79.18$_{\pm0.3}$	36.28$_{\pm0.5}$	13.48$_{\pm0.6}$	78.42$_{\pm0.3}$	57.23$_{\pm0.7}$	38.64$_{\pm0.8}$	62.54$_{\pm0.5}$	27.91$_{\pm0.8}$
BE	13.21$_{\pm0.4}$	76.74$_{\pm0.3}$	67.14$_{\pm0.5}$	-	**85.59**$_{\pm0.3}$	79.21$_{\pm0.6}$	14.98$_{\pm0.6}$	81.39$_{\pm0.4}$	**39.41**$_{\pm0.8}$	11.18$_{\pm0.8}$	85.66$_{\pm0.4}$	59.14$_{\pm0.7}$	42.61$_{\pm0.4}$	64.72$_{\pm0.3}$	28.96$_{\pm0.5}$
Poisoner	**3.57**$_{\pm0.3}$	**84.26**$_{\pm0.1}$	**78.01**$_{\pm0.4}$	-	84.81$_{\pm0.3}$	**83.02**$_{\pm0.5}$	**12.98**$_{\pm0.5}$	**82.57**$_{\pm0.2}$	37.40$_{\pm0.4}$	**8.77**$_{\pm0.3}$	**87.51**$_{\pm0.1}$	**64.75**$_{\pm0.2}$	**31.51**$_{\pm0.4}$	**66.28**$_{\pm0.3}$	**32.43**$_{\pm0.7}$
GroupDRO*	5.23$_{\pm0.2}$	86.72$_{\pm0.1}$	79.83$_{\pm0.2}$	-	81.53$_{\pm0.4}$	68.47$_{\pm0.5}$	16.51$_{\pm0.5}$	83.13$_{\pm0.4}$	29.61$_{\pm0.7}$	10.58$_{\pm0.4}$	85.03$_{\pm0.2}$	42.70$_{\pm0.4}$	29.22$_{\pm0.5}$	67.76$_{\pm0.3}$	44.32$_{\pm0.7}$

C STANDARD DEVIATION OF MAIN PAPER

Due to space limitations in the main paper, we did not report the standard deviation. In this appendix, we supplement the experimental results of the main paper by providing the standard deviation in Table 2 and Table 3.