Unsupervised Learning of Debiased Representations with Pseudo-Attributes
Supplementary Document

A. Full experimental results

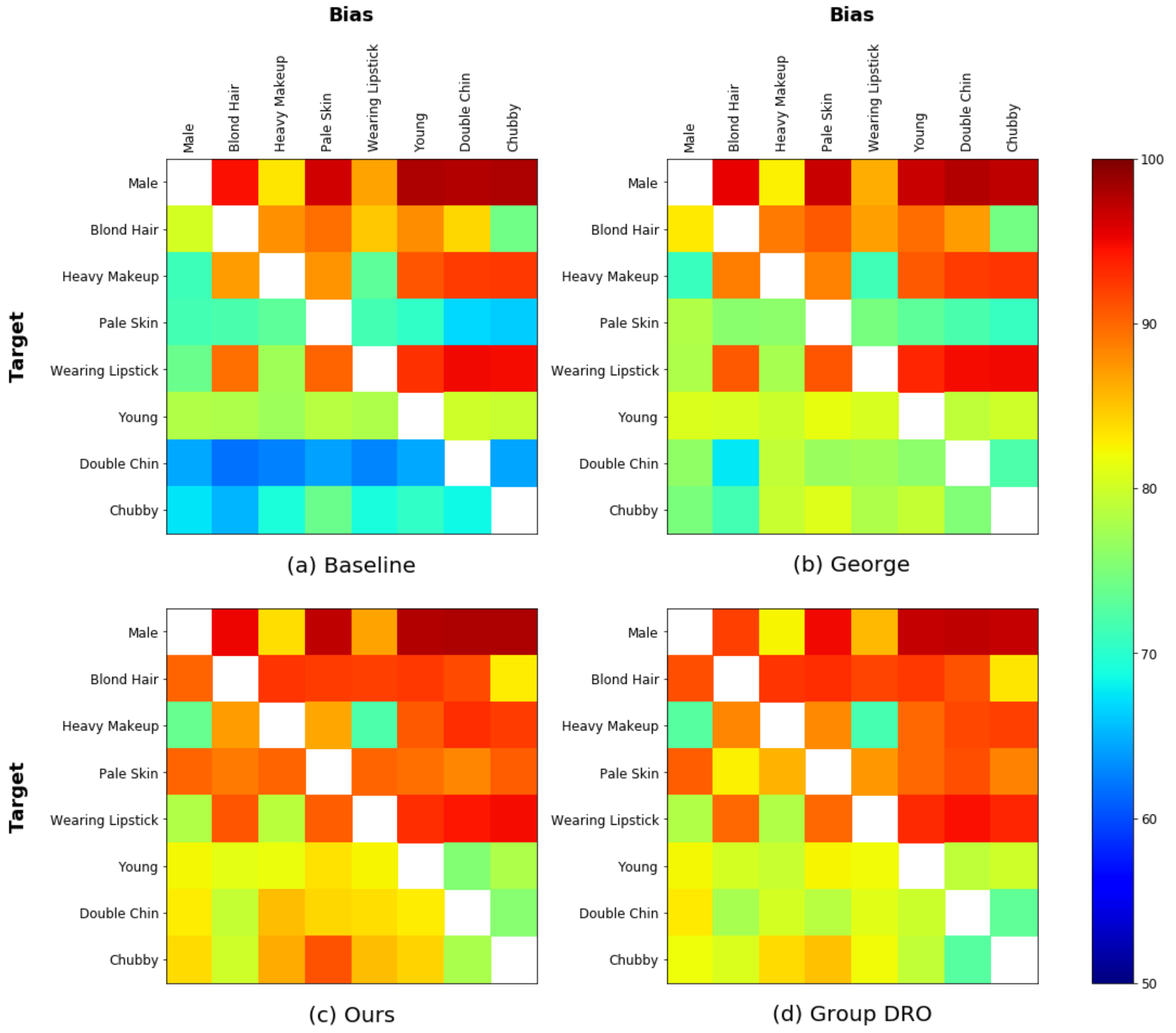
Table A and B present the full results of Table 1 in the main paper, including George [3] and a class weighting method. George [3] is closely related to our ablative model with sample weighting based on its loss, which is shown in Table 6 of the main paper, while class weighting approach adjusts the weight of each example depending on the associated class scale (size) to mitigate the class imbalance issue. We also report the gap between the overall accuracy and the unbiased accuracy of the baseline model to present the degree of algorithmic bias for each target attribute with *gender* bias. Bold and underline fonts indicate the first and second place among the compared approaches, respectively. The proposed approach achieves outstanding performance compared to all other unsupervised methods, and is even as competitive as the supervised counterpart [2]. Also, it is surprising that the class weighting method is superior to existing unsupervised debiasing methods including LfF [1] and George [3]. We run all experimental three times and compute average accuracies and their standard deviations.

Table A. Unbiased accuracy (%) in the presence of spurious correlations between target and bias attributes on the test split of the CelebA dataset.

					Unsupervised			Supervised
Target	Gap (%p)	Overall	Baseline	LfF* [1]	George [3]	Class weighting	**Ours**	Group DRO [2]
Blond Hair	-15.28	95.70	80.42 ± 0.51	84.89 ± 0.14	83.13 ± 1.86	83.35 ± 0.85	90.18 ± 0.23	**91.39** ± 0.27
Heavy Makeup	-19.63	90.82	71.19 ± 0.37	71.85 ± 0.17	70.91 ± 0.77	71.74 ± 0.83	**73.78** ± **0.25**	72.70 ± 0.71
Pale Skin	-25.25	96.75	71.50 ± 1.60	75.23 ± 0.74	78.22 ± 3.75	90.02 ± 0.56	90.06 ± 0.75	**90.55** ± **0.84**
Wearing Lipstick	-18.70	92.60	73.90 ± 0.53	73.84 ± 0.05	78.05 ± 0.98	72.89 ± 1.28	**78.28** ± **0.88**	78.26 ± 2.73
Young	-9.30	87.49	78.19 ± 0.39	79.58 ± 0.14	80.79 ± 0.20	82.13 ± 0.82	82.27 ± 0.65	**82.40** ± **0.48**
Double Chin	-31.32	95.93	64.61 ± 0.82	68.47 ± 0.22	76.23 ± 0.11	82.13 ± 1.43	82.92 ± 0.54	**83.19** ± **1.11**
Chubby	-27.97	95.39	67.42 ± 0.95	71.56 ± 0.52	74.88 ± 1.91	79.64 ± 0.56	**83.88** ± **0.36**	81.90 ± 0.20
Wearing Hat	-5.57	99.10	93.53 ± 0.37	94.81 ± 0.15	95.72 ± 0.71	96.16 ± 0.50	96.80 ± 0.26	**96.84** ± **0.46**
Oval Face	-10.40	73.10	62.70 ± 0.62	62.30 ± 0.21	65.16 ± 0.23	65.13 ± 1.05	**67.18** ± **0.82**	65.40 ± 0.14
Pointy Nose	-11.81	73.91	62.10 ± 0.74	63.83 ± 0.28	61.68 ± 1.59	66.82 ± 2.76	68.90 ± 0.90	**70.71** ± **0.28**
Straight Hair	-12.24	82.52	70.28 ± 1.06	72.84 ± 0.12	77.80 ± 0.19	77.46 ± 0.70	**79.18** ± **0.38**	77.04 ± 0.70
Blurry	-22.98	96.03	73.05 ± 1.28	77.52 ± 0.45	81.28 ± 0.28	87.75 ± 0.87	**88.93** ± **0.32**	87.05 ± 0.90
Narrow Eyes	-23.29	86.47	63.18 ± 1.05	67.77 ± 0.08	68.03 ± 0.11	70.99 ± 0.60	76.39 ± 0.64	**76.72** ± **1.98**
Arched Eyebrows	-12.09	81.81	69.72 ± 0.37	71.87 ± 0.10	73.25 ± 0.29	75.58 ± 1.13	74.77 ± 0.69	**78.30** ± **1.79**
Bags Under Eyes	-14.16	83.63	69.47 ± 0.57	71.86 ± 0.05	74.81 ± 0.38	76.36 ± 1.05	**77.84** ± **1.14**	75.88 ± 1.18
Bangs	-6.37	95.41	89.04 ± 0.47	89.04 ± 0.50	92.62 ± 0.12	93.09 ± 0.29	93.94 ± 0.57	**94.45** ± **0.17**
Big Lips	-8.99	69.86	60.87 ± 0.58	62.15 ± 0.06	64.99 ± 0.13	63.74 ± 0.56	**66.50** ± **0.24**	63.70 ± 0.44
No Beard	-22.73	95.84	73.11 ± 0.90	73.13 ± 0.89	77.90 ± 0.20	77.83 ± 2.29	**79.58** ± **0.14**	77.86 ± 1.35
Receding Hairline	-23.31	93.03	69.72 ± 0.78	74.58 ± 0.21	78.86 ± 0.40	82.97 ± 0.97	84.95 ± 0.49	**85.15** ± **1.31**
Wavy Hair	-9.19	82.29	73.10 ± 0.56	74.53 ± 0.17	77.39 ± 0.15	76.50 ± 0.65	**79.89** ± **0.71**	79.65 ± 0.63
Wearing Earrings	-17.18	89.35	72.17 ± 0.91	74.17 ± 0.33	80.65 ± 0.04	78.65 ± 0.28	**84.57** ± **0.69**	83.50 ± 0.63
Wearing Necklace	-30.73	85.77	55.04 ± 0.59	57.21 ± 0.76	58.79 ± 0.10	67.05 ± 1.37	**68.96** ± **0.12**	62.89 ± 3.69
Big Nose	-14.74	82.44	67.70 ± 1.11	69.75 ± 0.03	71.85 ± 0.18	70.52 ± 1.02	**74.21** ± **0.43**	73.73 ± 0.27
Brown Hair	-8.88	86.95	78.07 ± 0.87	78.93 ± 1.24	83.07 ± 0.07	83.12 ± 0.38	83.83 ± 0.66	**84.87** ± **0.07**
Bushy Eyebrows	-17.02	91.44	74.42 ± 0.91	75.20 ± 0.34	80.99 ± 0.32	82.73 ± 1.21	85.02 ± 0.02	**85.43** ± **0.19**
Gray Hair	-20.54	98.01	77.47 ± 0.67	80.09 ± 0.21	86.10 ± 1.18	90.12 ± 1.12	91.80 ± 0.22	**92.52** ± **0.14**
Average	-16.91	88.52	71.61	73.73	76.66	78.63	**80.93**	80.46

Table B. Worst-group accuracy (%) in the presence of spurious correlation between target and bias attributes on the test split of the CelebA dataset.

					Unsupervised			Supervised
Target	Gap (%p)	Overall	Baseline	LfF* [1]	George [3]	Class weighting	**Ours**	Group DRO [2]
Blond Hair	-54.68	95.70	41.02 ± 1.96	57.96 ± 2.00	65.45 ± 15.52	53.58 ± 3.10	82.54 ± 1.22	**87.86 ± 0.10**
Heavy Makeup	-73.47	90.82	17.35 ± 4.60	23.87 ± 2.79	9.09 ± 1.24	28.86 ± 11.91	**39.84 ± 2.28**	21.36 ± 1.36
Pale Skin	-60.11	96.75	36.64 ± 3.53	43.26 ± 1.40	62.03 ± 16.50	85.42 ± 1.70	**88.60 ± 1.48**	87.68 ± 2.37
Wearing Lipstick	-61.22	92.60	31.38 ± 4.27	31.92 ± 0.02	51.04 ± 2.59	27.68 ± 3.45	**46.52 ± 1.62**	46.08 ± 5.57
Young	-34.70	87.49	52.79 ± 1.45	57.79 ± 0.84	65.12 ± 0.88	71.43 ± 1.75	74.33 ± 0.70	**76.29 ± 1.96**
Double Chin	-74.60	95.93	21.33 ± 2.24	28.24 ± 0.46	50.00 ± 0.41	62.43 ± 4.71	67.78 ± 0.91	**72.94 ± 1.14**
Chubby	-71.09	95.39	24.30 ± 3.73	34.09 ± 0.90	58.01 ± 11.04	52.76 ± 2.59	72.32 ± 0.93	**72.64 ± 1.70**
Wearing Hat	-13.98	99.10	85.12 ± 0.31	88.31 ± 0.12	92.93 ± 0.76	93.61 ± 0.32	**94.94 ± 0.19**	94.67 ± 0.41
Oval Face	-43.95	73.10	29.15 ± 2.76	36.00 ± 1.46	38.01 ± 2.63	43.52 ± 6.37	55.78 ± 0.94	**56.84 ± 1.83**
Pointy Nose	-48.11	73.91	25.80 ± 4.03	38.04 ± 1.49	22.63 ± 3.67	47.46 ± 3.75	52.48 ± 0.52	**63.76 ± 2.80**
Straight Hair	-34.70	82.52	47.82 ± 6.75	58.53 ± 1.61	69.23 ± 1.24	68.97 ± 1.15	**72.09 ± 0.76**	66.10 ± 3.56
Blurry	-50.35	96.03	45.68 ± 3.98	52.35 ± 1.18	62.23 ± 1.58	82.30 ± 3.05	**84.10 ± 0.73**	82.06 ± 2.27
Narrow Eyes	-59.46	86.47	27.01 ± 1.30	38.53 ± 0.44	35.16 ± 1.14	52.62 ± 4.11	**73.24 ± 0.88**	71.47 ± 3.72
Arched Eyebrows	-47.05	81.81	34.76 ± 1.86	44.97 ± 0.46	45.64 ± 1.21	52.94± 5.28	54.36 ± 1.37	**69.44 ± 5.44**
Bags Under Eyes	-41.98	83.63	41.65 ± 1.01	49.10 ± 0.49	56.28 ± 2.11	59.77 ± 8.13	62.55 ± 0.90	**63.34 ± 3.02**
Bangs	-18.50	95.41	76.91 ± 3.27	82.37 ± 0.52	85.90 ± 0.24	87.91 ± 1.80	**92.21 ± 1.24**	92.12 ± 1.03
Big Lips	-39.01	69.86	30.85 ± 0.62	38.54 ± 0.18	44.51 ± 0.83	43.16 ± 5.62	**56.99 ± 3.05**	47.55 ± 1.03
No Beard	-82.54	95.84	13.30 ± 3.87	20.00 ± 0.00	33.33 ± 5.77	30.00 ± 10.00	**40.00 ± 0.00**	36.70 ± 5.10
Receding Hairline	-57.34	93.03	35.69 ± 0.35	45.53 ± 0.55	57.30 ± 0.90	72.14 ± 2.56	**79.12 ± 1.91**	**79.12 ± 2.11**
Wavy Hair	-44.28	82.29	38.01 ± 0.85	45.24 ± 0.83	53.17 ± 0.43	49.69 ± 4.65	65.74 ± 1.13	**66.79 ± 1.62**
Wearing Earrings	-63.09	89.35	26.26 ± 4.14	32.95 ± 1.31	52.74 ± 1.10	47.18 ± 4.08	72.81 ± 1.50	**75.24 ± 2.10**
Wearing Necklace	-83.05	85.77	2.72 ± 0.83	6.67 ± 2.07	13.82 ± 0.41	30.36 ± 3.36	**41.93 ± 2.47**	24.34 ± 7.81
Big Nose	-49.25	82.44	33.19 ± 3.97	45.30 ± 0.50	46.22 ± 0.41	49.56 ± 4.79	63.00 ± 4.27	**65.08 ± 1.17**
Brown Hair	-27.37	86.95	59.58 ± 2.55	60.68 ± 3.62	73.20 ± 0.88	70.91 ± 3.09	71.50 ± 0.97	**78.92 ± 1.61**
Bushy Eyebrows	-54.30	91.44	37.14 ± 2.54	52.67 ± 3.14	56.08 ± 0.97	66.92 ± 6.98	74.08 ± 0.75	**81.56 ± 3.24**
Gray Hair	-55.52	98.01	42.49 ± 1.86	48.46 ± 1.09	67.23 ± 2.75	80.00 ± 3.78	83.03 ± 1.37	**88.55 ± 1.85**
Average	51.68	88.79	36.84	44.67	50.39	58.00	67.76	**68.02**



(a) Baseline

(b) George

(c) Ours

(d) Group DRO

Figure A. Heatmap of unbiased accuracy (%) with 4 different methods. Unlike previous tables, we evaluate our model with various bias attributes, in addition to *Male* (gender), on the CelebA dataset. To be specific, we select 8 attributes and evaluate unbiased accuracies with all possible (target, bias) pairs among the attributes. For each figure, the columns and rows denote bias and target attributes, respectively. Our approach substantially improves unbiased accuracies for various bias attributes consistently.

B. Additional Analysis

Unbiased results with various bias attributes To make our study more comprehensive, we also evaluate our model with various bias attributes, in addition to *Male* (gender), on the CelebA dataset. Specifically, we select 8 attributes[1] and test our model with all possible (target, bias) pairs among the attributes. Figure A visualizes the experimental results with different methods, including baseline, George [3], group DRO [2] and our approach, in terms of unbiased accuracy (%). The columns and rows denote bias and target attributes, respectively. As shown in the figure, our model improves unbiased accuracies substantially for various bias attributes, which outperforms baseline and George [3] and is even as competitive as group DRO [2].

Algorithmic bias with various bias attributes In Figure B, we visualize the performance gap between overall accuracy and unbiased accuracy for each method to analyze the degree of algorithmic bias between target and bias attributes. We use

[1] The selected attributes are male, blond hair, heavy makeup, pale skin, wearing lipstick, young, double chin and chubby.

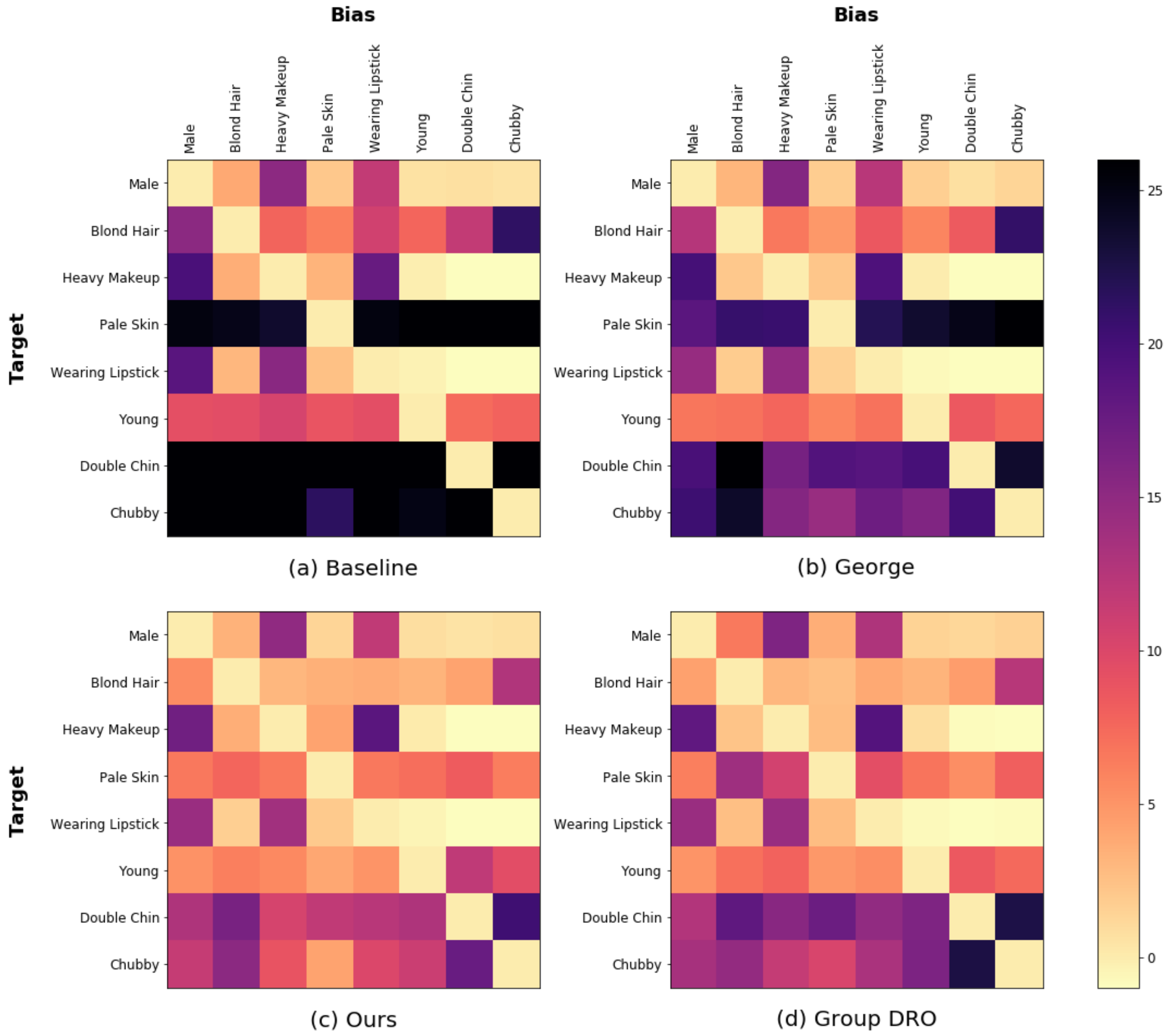


Figure B. Heatmap of the performance gap between overall accuracy and unbiased accuracy (%p) with 4 different methods. We use the same experimental setup with Figure A. The columns and rows denote bias and target attributes, respectively. In subfigure (a), the larger the performance gap, the more severe the algorithmic bias. As shown in the figure, even for the same target attribute, the gap varies largely depending on bias attributes. Subfigure (b), (c) and (d) demonstrate that all methods mitigate the algorithmic bias, while our approach is more effective than George.

the same experimental setting with Figure A. The larger the performance gap, the more severe the algorithmic bias. This implies that, based on the performance gap from Figure B (a), we can measure the existence of algorithmic bias on the CelebA dataset, *e.g.*, the target attribute *Heavy Makeup* is spuriously correlated to *Male* and *Wearing Lipstick* biases while not to *Young*, *Double Chin* and *Chubby* biases.[2] As shown in the figure, even with the same target attribute, the gap varies largely depending on bias attributes. We also observe that the algorithmic bias does not exist symmetrically, *e.g.*, the target attribute *Chubby* is spuriously correlated to *Heavy Makeup* bias, not vice versa. Compared to the baseline, all methods reduce the algorithmic bias while our framework is more effective than George [3] and as competitive as group DRO [2].

Multi-target classification We tested our framework with another setting, called multi-target classification, where a single backbone model adopts multiple classification heads. To this end, we attached multiple linear classification layers, which

[2] As in the main paper, we suppose that the algorithmic bias exists between target and bias attributes when a baseline model gives a large performance gap between its overall accuracy and unbiased accuracy (e.g., > 5% points).

Table C. Unbiased accuracy (%) with multi-target classification scenario. In this setting, each model is trained to classify multiple attributes jointly by adopting separate linear branches. The bias attribute is fixed to *Male*. We report the unbiased accuracy for each target attribute separately.

		Unsupervised		Supervised
Targets	Baseline	George [3]	Ours	Group DRO [2]
Blond Hair / Heavy Makeup	78.92 / 71.46	83.06 / 70.99	89.78 / **72.25**	**90.38** / 70.94
Blond Hair / Wearing Lipstick	80.76 / 71.91	82.08 / 73.06	**89.09** / 77.34	88.86 / **78.45**
Straight Hair / Oval Face	69.93 / 60.84	76.77 / 63.84	**78.33 / 64.85**	76.38 / 64.77
Straight Hair / Big Lips	70.05 / 60.14	69.73 / 63.22	76.03 / **66.39**	**77.06** / 64.07
Blurry / Pale Skin	73.89 / 68.18	79.51 / 79.45	87.35 / **89.28**	**87.82** / 85.91
Blurry / Young	76.48 / 77.82	74.66 / 77.16	**88.64 / 82.79**	88.57 / 82.14

correspond to individual targets, respectively, to a shared feature extractor. For evaluation, we calculate unbiased accuracy for each target attribute separately, where the bias attribute is fixed to *gender*. Table C presents the multi-target classification results with several target attribute pairs, where our model achieves consistently better results than the compared unsupervised method in terms of unbiased accuracy, while it is as competitive as group DRO [2].

References

[1] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *NeurIPS*, 2020. 1, 2

[2] 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. In *ICLR*, 2020. 1, 2, 3, 4, 5

[3] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS*, 2020. 1, 2, 3, 4, 5