A APPENDIX

A.1 REPLICATION EXPERIMENTS

Single-Attribute Replication Experiments. We replicate single attribute experiment settings considered in previous works and introduce more competitive baselines. We find, as shown in Tables 7, 8 and 9, that Weighted ERM is a consistently competitive baseline, even outperforming methods in Wang et al. (2020). We also confirm prior findings that methods not needing attribute labels, i.e. unlabeled methods (SD, JTT), perform almost as well as ones that use attribute labels (Liu et al., 2021; Shrestha et al., 2021).

Table 7: Shrestha et al. (2021) Replication (CelebA-SL)

Model	Inter. Bias	Bias Amp	Accuracy	Reweighted Accuracy	Min Accuracy
ERM	1.005 ± 0.109	-0.024 ± 0.012	0.923 ± 0.012	0.848 ± 0.012	0.604 ± 0.065
GDRO	0.558 ± 0.049	-0.065 ± 0.008	0.911 ± 0.010	0.895 ± 0.009	0.833 ± 0.034
IRMv1	1.071 ± 0.148	-0.010 ± 0.010	0.940 ± 0.007	0.793 ± 0.009	0.404 ± 0.013
SD	0.746 ± 0.032	-0.057 ± 0.004	0.884 ± 0.005	0.887 ± 0.005	0.817 ± 0.025
IRMv1 Batch	0.603 ± 0.019	-0.061 ± 0.004	0.911 ± 0.008	0.891 ± 0.009	0.815 ± 0.039
Weighted ERM	0.610 ± 0.069	-0.059 ± 0.008	0.915 ± 0.006	0.903 ± 0.007	0.852 ± 0.041

Table 8: Liu et al. (2021) Replication (CelebA-SL)

Model	Inter. Bias	Bias Amp	Accuracy	Reweighted Accuracy	Min Accuracy
ERM	1.107 ± 0.606	0.024 ± 0.018	0.930 ± 0.044	0.715 ± 0.148	0.281 ± 0.197
GDRO	0.666 ± 0.010	-0.049 ± 0.002	0.936 ± 0.002	0.916 ± 0.001	0.857 ± 0.003
JTT	0.911 ± 0.005	-0.041 ± 0.002	0.902 ± 0.005	0.907 ± 0.001	0.842 ± 0.003
Weighted ERM	0.647 ± 0.101	-0.055 ± 0.012	0.922 ± 0.010	0.921 ± 0.003	0.891 ± 0.026

Table 9: Wang et al. (2020) Replication (CelebA-ML)

Model	mAP ↑	Reweighted mAP ↑	Inter. Bias ↓	Bias Amp ↓
ERM	0.794 ± 0.001	0.746 ± 0.001	1.179 ± 0.037	0.007 ± 0.002
Discriminative	0.792 ± 0.001	0.739 ± 0.002	1.176 ± 0.012	0.007 ± 0.003
GDRO	0.639 ± 0.003	0.569 ± 0.002	1.316 ± 0.018	0.039 ± 0.003
Independent	0.780 ± 0.001	0.760 ± 0.000	0.854 ± 0.021	-0.029 ± 0.004
Independent-SP	0.779 ± 0.003	0.757 ± 0.002	0.837 ± 0.026	-0.025 ± 0.005
Adversarial	0.770 ± 0.001	0.706 ± 0.002	1.324 ± 0.026	0.026 ± 0.002
Weighted ERM	0.767 ± 0.001	0.772 ± 0.004	0.774 ± 0.007	-0.061 ± 0.003

A.2 EXPERIMENTS WITH DIFFERENT PROTECTED ATTRIBUTES

Our results on intersectional bias in Figures 2 and 3 assume a fixed set of protected attributes. This raises the question of whether our findings would be different if the protected attributes were differently selected—for instance, what if the protected attributes were more noisily labeled or had especially balanced/imbalanced label distributions? In Figures 4, 5, and 6, we repeat the experiments illustrated in Figures 2 and 3 but for different choices of protected attributes. In particular, we consider sets of protected attributes composed of (1) protected attributes with balanced label distributions, (2) protected attributes with particularly imbalanced label distributions, and (3) protected attributes with noisy labels. For all three choices, we observe the same trends as in Figure 2 and 3, showing that the trends highlighted in our empirical discussion hold generally for different choices of protected attributes.

A.3 EXPERIMENT DETAILS

A.3.1 COMPUTING ENVIRONMENT

Each experiment run, corresponding to the training of a single model, is trained concurrently with two other runs on their own GPU type Tesla V100-SXM2-32GB-LS provisioned from a commercial cloud service. The training process for three such concurrent runs takes anywhere from two hours up to twenty four hours.

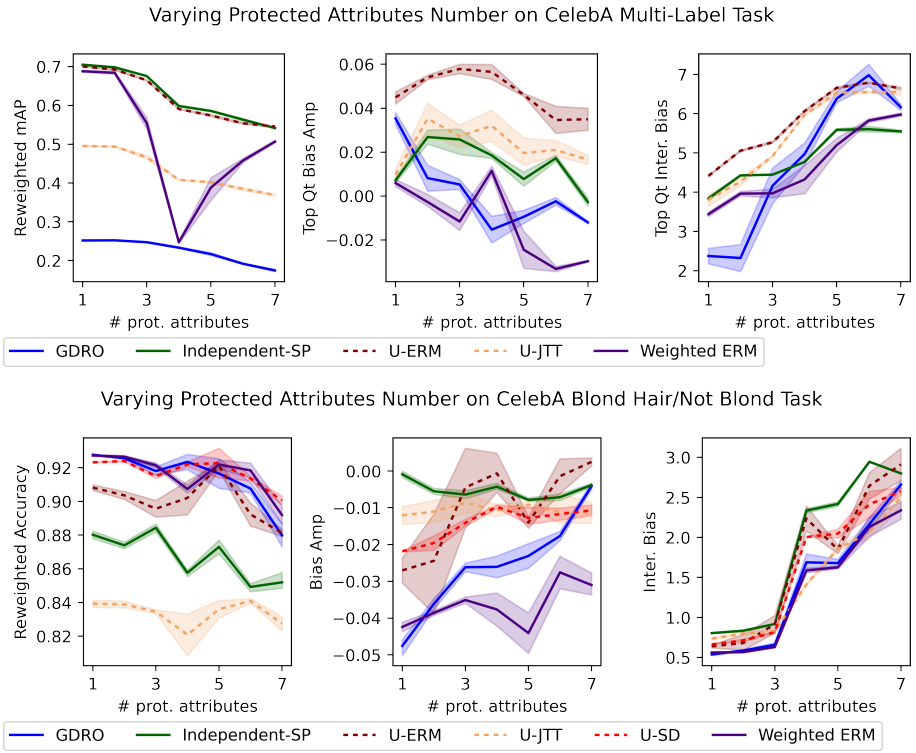

Figure 4: Bias mitigation algorithms on the CelebA-ML task (top) and CelebA-SL task (bottom) for $k \in [1, 7]$ protected attributes: Wearing Lipstick, High Cheekbones, Heavy Makeup, Male, Attractive, Smiling, Mouth Slightly Open. These protected attributes are, in order, the CelebA attributes with the most balanced labels (i.e. close to 50-50). The left figure plots the reweighted mean-average-precision/accuracy, the middle figure plots bias amplification, and the right figure plots intersectional bias.

A.3.2 Dataset Information

Here, we detail the five datasets that our experiments use in training time. The first four are variants of CelebA (Liu et al., 2015), a dataset of celebrity facial pictures.

The CelebA dataset provides a multi-label task of predicting 39 binary labels such as whether the pictured celebrity has "Narrow Eyes" and leaves the "is Male" label as protected. The CelebA-Multi dataset adds 6 other protected labels; during experiments on this dataset, we increment from one to seven protected labels—ignoring any of the seven which are not protected.

The CelebA-Class dataset provides a binary classification task of predicting "Blond Hair" labels and leaves the "is Male" label as protected. The CelebA-Class Multi dataset adds 6 other protected labels to CelebA-Class; during experiments on this dataset, we increment from one to seven protected labels—ignoring any of the seven which are not protected.

The ImageNet dataset we use is based on the People subtree of the ImageNet challenge (Deng et al., 2009). This is a multi-class task. We use protected attribute labels about gender, skin color and age provided by Yang et al. (2020); these attribute labels only cover around $15,000$ of the $140,000$ images we were able to download from the ImageNet people subtree.

The following tables detail additional dataset information.

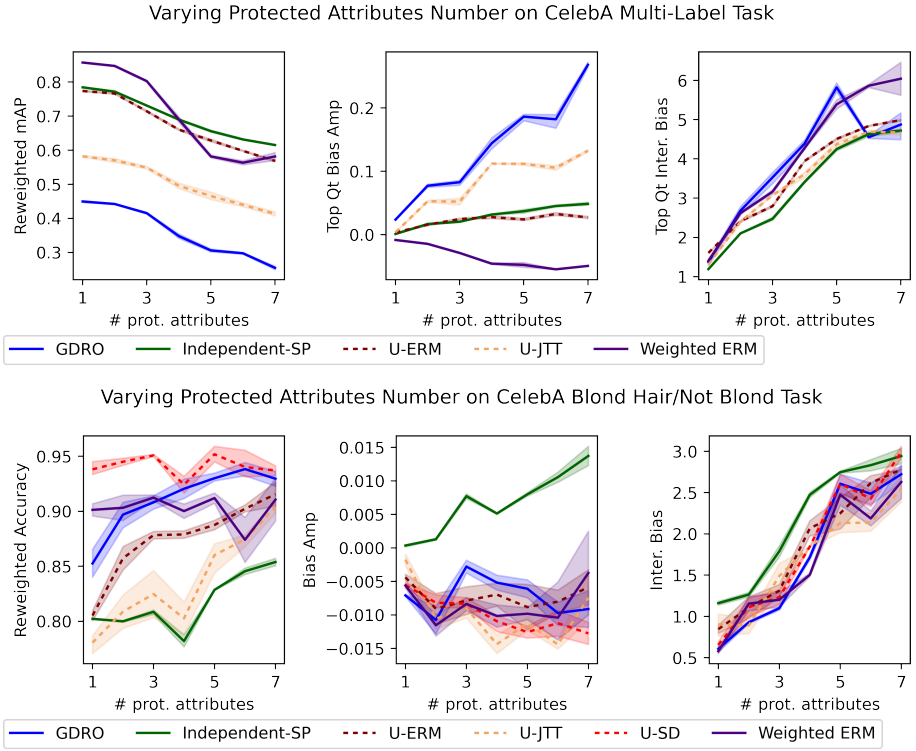

Figure 5: Bias mitigation algorithms on the CelebA-ML task (top) and CelebA-SL task (bottom) for $k \in [1, 7]$ protected attributes: Wearing Hat, Double Chin, Blurry, Gray Hair, Bald, Sideburns, Mustache. These protected attributes are, in order, the CelebA attributes with the most imbalanced labels (i.e. close to 50-50). The left figure plots the reweighted mean-average-precision/accuracy, the middle figure plots bias amplification, and the right figure plots intersectional bias.

	# Protected Groups	# Prediction Classes	Protected Attributes
CelebA Multi	$0 - 2^7$	2	Pale Skin, Male, Narrow Eyes, Big Nose, Young, Straight Hair, Attractive
CelebA-Class Multi	$0 - 2^7$	33	Pale Skin, Male, Narrow Eyes, Big Nose, Young, Straight Hair, Attractive
ImageNet	196	284	Gender, Skin Color, Age
CelebA	2	34	Male
CelebA-Class	2	2	Male

	Training Set Size	Attribute-Labeled Training Set Size	Eval Size	Test size
CelebA	162770	162770	19867	19962
CelebA Multi	162770	162770	19867	19962
CelebA-Class	162770	162770	19867	19962
CelebA-Class Multi	162770	162770	19867	19962
ImageNet	5861	124693	5327	5327

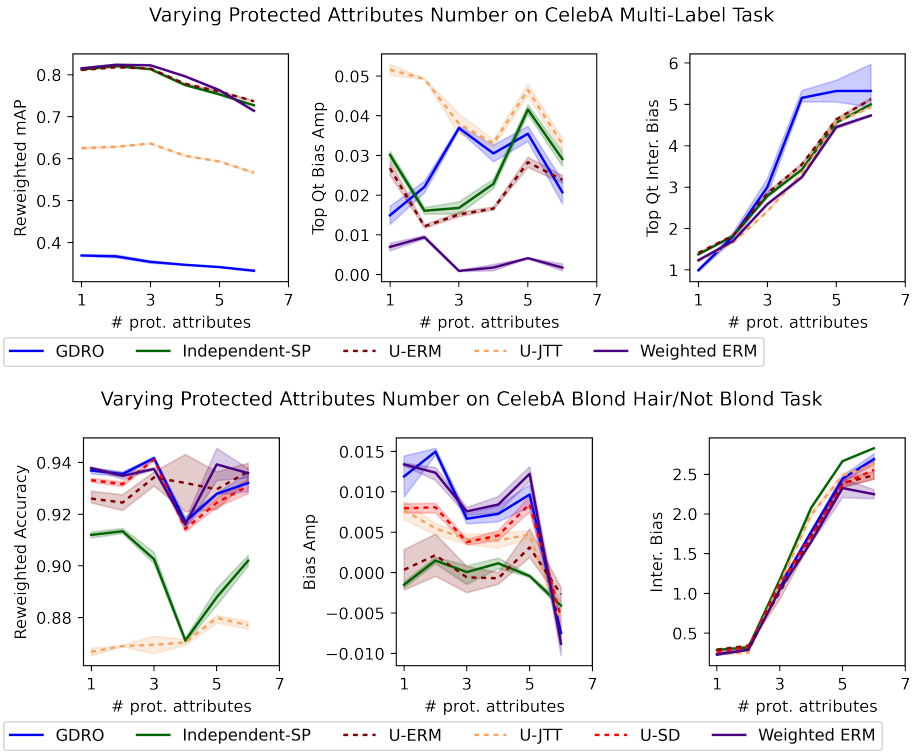

Figure 6: Bias mitigation algorithms on the CelebA-ML task (top) and CelebA-SL task (bottom) for $k \in [1, 7]$ protected attributes: Wavy Hair, Oval Face, Big Nose, Pale Skin, Big Lips, Straight Hair. These protected attributes are the CelebA attributes designated as "inconsistently labeled" by Ramaswamy et al. (2021). The left figure plots the reweighted mean-average-precision/accuracy, the middle figure plots bias amplification, and the right figure plots intersectional bias.

	Normalization	Augmentation
CelebA	By mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225], center cropped and resized to (224, 224)	Random resized crop of (224, 224) from (256, 256) and random horizontal flips
CelebA Multi	By mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225], center cropped and resized to (224, 224)	Random resized crop of (224, 224) from (256, 256) and random horizontal flips
CelebA-Class	By mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225], center cropped and resized to (224, 224)	None
CelebA-Class Multi	By mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225], center cropped and resized to (224, 224)	None
ImageNet	By mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225], center cropped and resized to (224, 224)	Random resized crop of (224, 224) from (256, 256) and random horizontal flips

A.3.3 Experiment Hyperparameters

Here, we detail the hyperparameters used in our experiments. Unless otherwise-specified, the hyperparameters listed in these tables were determined by a grid search over a combination of the values listed in the below table.

	Range
Learning Rate	1e-2, 1e-3, 1e-4, 1e-5
Batch Size	32, 128
Weight Decay	1e-1, 1e-4, 0
Dropout	0, 0.5
Group learning rates (Group DRO	1, 0.1, 0.01
Gradient penalty (IRM)	0.2, 1, 5
Groups sampled per batch (IRM)	1, 4, 16
Initial Epochs (Just Train Twice)	1, 5, 30
Importance weight (Just Train Twice)	1, 5, 20, 50

Table 8 is an experiment run on the CelebA Class dataset. All hyperparameters seen are chosen to match Liu et al. (2021) as closely as possible. As with the original paper, we use a Resnet-50 He et al. (2016) trained with SGD with momentum 0.9, pretrained on ImageNet. The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
ERM	1e-4	128	1e-4	0	50	N/A
Weighted	1e-4	128	1e-4	0	50	N/A
Group DRO	1e-5	128	1e-1	0	50	Group learning rate of 0.01
Just Train Twice	1e-5	128	1e-1	0	50	Importance weight of $\lambda = 50$, using ERM model trained for 1 epoch.

Table 7 is an experiment run on the CelebA Class dataset. All hyperparameters seen are chosen to match Shrestha et al. (2021) as closely as possible. As with the original paper, we use a Resnet-18 He et al. (2016) trained with SGD with momentum 0.9. The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
ERM	1e-3	128	0	0	50	N/A
Weighted	1e-5	128	1e-1	0	50	N/A
Group DRO	1e-5	128	1e-1	0	50	Group learning rate of 0.01
IRM	1e-4	128	0	0	50	Gradient penalty of 1
Spectral Decoupling	1e-4	128	1e-5	0	50	Per class $\lambda = (10, 10)$, $\gamma = (0.44, 0.25)$.

Table 9 is an experiment run on the CelebA dataset. All hyperparameters seen are chosen to match Wang et al. (2020) as closely as possible. As with the original paper, we use a Resnet-50 He et al. (2016) trained with Adam, pretrained on ImageNet. The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
ERM	1e-4	128	0	0.5	50	N/A
Weighted	1e-4	128	0	0.5	50	N/A
Independent	1e-4	128	0	0.5	50	N/A
Independent SP	1e-4	128	0	0.5	50	N/A
Discriminative	1e-4	128	0	0.5	50	N/A
Adversarial	1e-4	128	0	0.5	50	Training ratio (adversarial:main) 3:1, confusion loss weight $= 1.0$

Figures 3,4, 5, and 6 are experiments run on the CelebA Multi dataset. Hyperparameters were chosen by grid search. We use a Resnet-50 He et al. (2016) trained with Adam, pretrained on ImageNet—the same settings as Wang et al. (2020). The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
ERM	1e-4	32	0	0.5	30	N/A
Weighted	1e-4	32	0	0.5	30	N/A
Independent SP	1e-4	32	0	0.5	30	N/A
Group DRO	1e-4	32	0	0.5	30	Group learning rate of 0.1
Just Train Twice	1e-4	32	0	0.5	30	Importance weight $\lambda = 20$, using ERM model trained for 1 epoch

Figures 2, **??**, **??**, and **??** are experiments run on the CelebA-Class Multi dataset. Hyperparameters were chosen by grid search. We use a Resnet-50 He et al. (2016) trained with SGD with momentum 0.9, pretrained on ImageNet—the same settings as Liu et al. (2021). The hyperparameters are listed below.

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
ERM	1e-4	128	1e-1	0	50	N/A
Weighted	1e-5	128	1e-1	0	50	N/A
Independent SP	1e-4	128	1e-4	0	50	N/A
Group DRO	1e-4	128	1e-1	0	50	Group learning rate of 0.01
Just Train Twice	1e-4	32	1e-1	0	50	Importance weight $\lambda = 5$, using ERM model trained for 1 epoch
IRM	1e-4	32	1e-1	0	50	Gradient penalty of 1
Uniform IRM	1e-4	128	1e-1	0	50	Gradient penalty of 1, 16 groups sampled per batch
Spectral Decoupling	1e-4	128	1e-4	0	50	Per class $\lambda = (10, 10)$, $\gamma = (0.44, 2.5)$

Tables 6, 4, 5 are experiments run on the ImageNet dataset. Hyperparameters were chosen by grid search. We use a Resnet-50 He et al. (2016) trained with SGD with momentum 0.9, with standard ImageNet pretrained weights—note that the ImageNet subset used for pretraining does not intersect with the ImageNet People Subtree we train on.[1] When pretraining on our data splits without protected attribute labels, all methods are initialized from the trained ERM models' weights (for the corresponding seed). The hyperparameters are listed below.

[1]Further note that while we initialize our network with standard ImageNet pretrained weights (trained on a different subset of ImageNet than we use), some of our experiments involve also pretraining on a subset of ImageNet that we do use (see Table 5).

	Learning Rate	Batch Size	Weight Decay	Dropout	Epochs	Custom Parameters
ERM	1e-4	64	1e-4	0	100	N/A
Weighted	1e-4	64	1e-2	0	50	N/A
Sqrt-Weighted	1e-4	64	1e-2	0	50	N/A
Sqrt-Weighted Distilled	1e-4	64	1e-2	0	50	Distill weight = 1.0
Group DRO	1e-4	64	1e-2	0	50	Group learning rate of 0.01
Just Train Twice	1e-5	64	1e-1	0	50	Importance weight of $\lambda = 5$, using ERM model trained for 5 epoch.
Independent SP	1e-4	64	1e-1	0	50	N/A
Independent SP Distilled	1e-4	64	1e-1	0	50	Distill weight = 1.0
Spectral Decoupling	1e-5	64	1e-1	0	50	Per class $\lambda = 10$, $\gamma = 0$