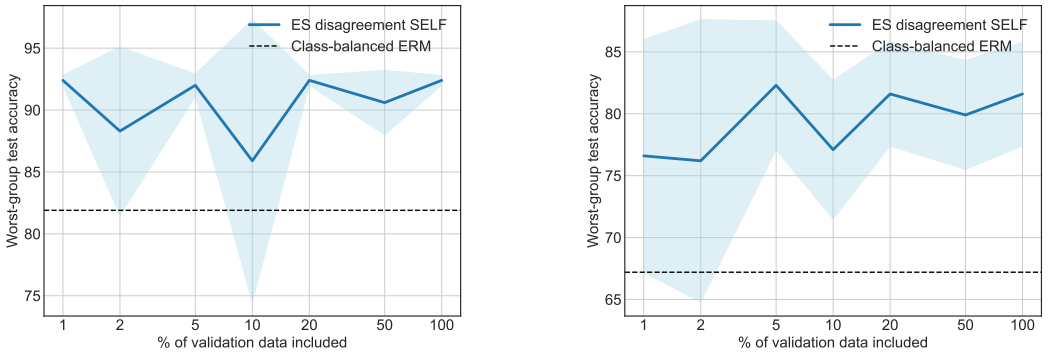
Submission 8174 Rebuttal Figures



(a) Waterbirds (b) CelebA

Figure 1: **Ablation on the size of the validation set.** We perform model selection with early-stop disagreement SELF using varying amounts of group-labeled validation data. The results show that ES disagreement SELF is robust to hyperparameter tuning and can massively reduce the annotation requirement: at 1% of data, Waterbirds has only 6 examples and CelebA has only 99 examples. The MultiNLI and CivilComments ablations will be done by next week and made available to the reviewers upon request. We plot the mean and standard deviation over three independent seeds.

Table 1: **Average accuracy performance**. We detail the average test accuracy of our methods on the 4 benchmark datasets. Both of our methods have similar average accuracy to DFR, which experiences a slight accuracy/robustness tradeoff compared to ERM (as is typical in the robustness literature). We list the mean and standard deviation over three independent seeds.

Method	Group Anns	Waterbirds	CelebA	CivilComments	MultiNLI
ERM	✗	$90.2_{\pm 0.7}$	$94.4_{\pm 0.2}$	$92.0_{\pm 0.2}$	$81.8_{\pm 0.2}$
CB last-layer retraining	✗	$94.8_{\pm 0.3}$	$93.6_{\pm 0.2}$	$87.1_{\pm 0.0}$	$82.0_{\pm 0.2}$
ES disagreement SELF	✗	$94.5_{\pm 0.6}$	$91.4_{\pm 0.4}$	$87.6_{\pm 0.8}$	$81.1_{\pm 1.1}$
DFR (our impl.)	✓	$94.9_{\pm 0.3}$	$92.6_{\pm 0.5}$	$87.5_{\pm 0.2}$	$81.7_{\pm 0.2}$

(a) Misclassification (b) Dropout disagreement (c) Early-stop disagreement

Figure 2: **Qualitative examples.** We display the top 4 CelebA datapoints with the highest loss or KL divergence for misclassification and disagreement methods, respectively. We use the model which achieved the highest validation worst-group accuracy to generate these images.