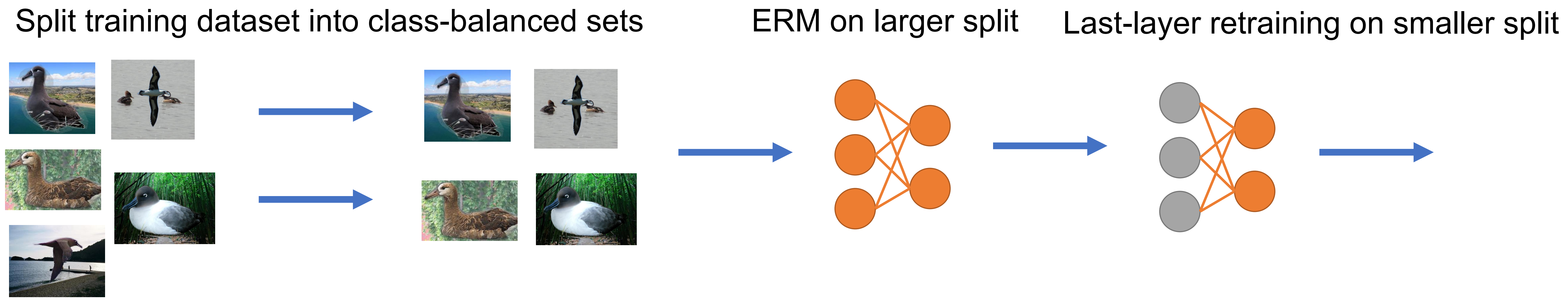
# Saving a Split for Last-Layer Retraining can Improve Group Robustness without Group Annotations

Tyler LaBonte[1], Vidya Muthukumar[1], Abhishek Kumar[2]
[1]Georgia Tech, [2]Google DeepMind

**Abstract:** Worst-group accuracy can be improved with no group annotations by holding out a split of the training dataset for class-balanced last-layer retraining.



Split training dataset into class-balanced sets

ERM on larger split

Last-layer retraining on smaller split

## Problem: Empirical risk minimization gives poor minority group performance

- Datasets often suffer from *spurious correlations* which are irrelevant for the true label
- Spurious features create minority groups which are underrepresented during training
- Maximize worst-group test accuracy (WGA) instead of mean over the training distribution (ERM)

Landbird on land (73%)

Landbird on water (4%)

Waterbird on water (22%)

Waterbird on land (1%)

## Prior Work: With group annotations, last-layer retraining greatly improves WGA

- Models learn core features, but spurious features are overweighted in last layer [1]
- Last-layer retraining (DFR) on held-out group-balanced dataset is efficient and effective
- However, groups are often unknown ahead of time or are difficult to annotate

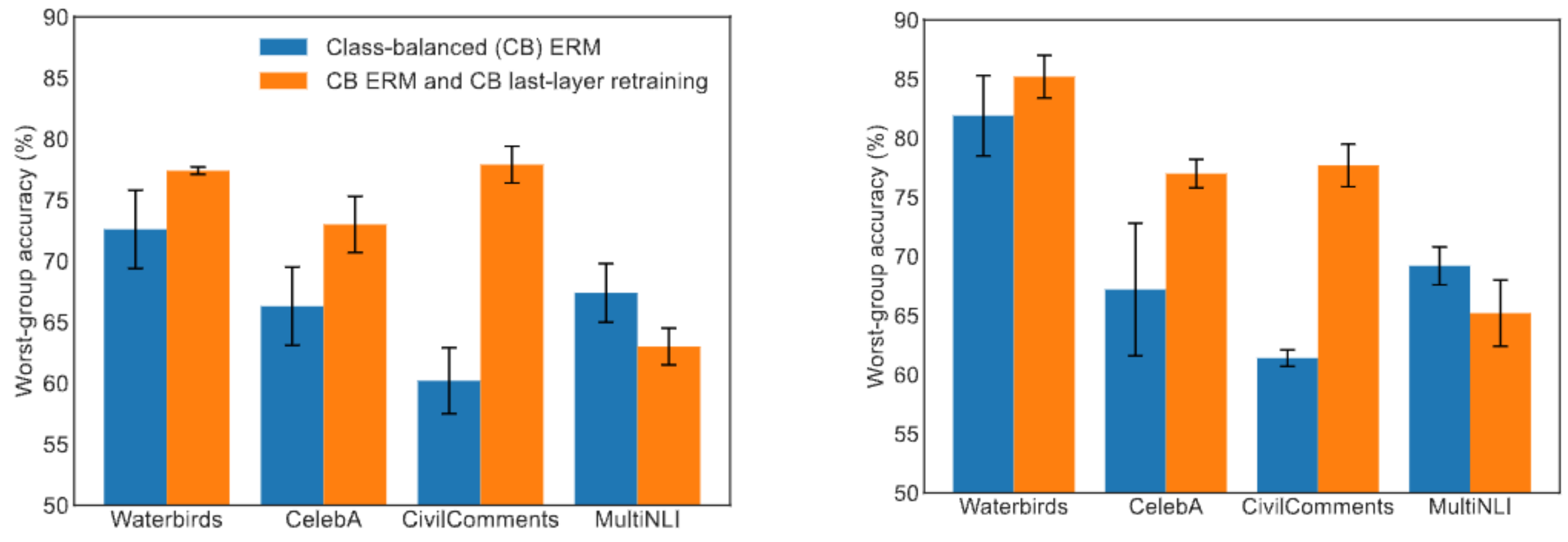
Class-balanced last-layer retraining on DFR held-out set, averaged over 3 random seeds.

| Method | Group Annotations | Worst-Group Test Accuracy | | | |
|---|---|---|---|---|---|
| | | Waterbirds | CelebA | CivilComments | MultiNLI |
| CB ERM | ✗ | 81.9 | 67.2 | 61.4 | **69.2** |
| JTT [2] | ✗ | 85.6 | 75.6 | 67.8 | 67.5 |
| RWY-ES [3] | ✗ | 74.5 | **76.8** | 78.9 | 68.0 |
| CB retraining (ours) | ✗ | **92.6** | 73.7 | **80.4** | 64.7 |
| DFR [1] | ✓ | 92.4 | 87.0 | 81.8 | 70.8 |

## Our Work: Class-balanced retraining on held-out data boosts WGA without group anns

- Significant DFR gain is due solely to class-balancing; achieves 94% of DFR acc on average
- Class-balanced last-layer retraining renders class-balancing in the ERM stage optional [3]
- Improves WGA despite similarly imbalanced groups in held-out set: surprising and unexplained
- Worse on MultiNLI dataset because ERM is not saturated (collect more data instead)

Last-layer retraining is a “free lunch” in group robustness on 3 of 4 benchmark datasets



(a) Training dataset only

(b) Combined training and held-out datasets

### References

**[1]** Kirichenko et al. “Last layer re-training is sufficient for robustness to spurious correlations.” NeurIPS 2022. **[2]** Liu et al. “Just train twice: Improving group robustness without training group information.” ICML 2021. **[3]** Idrissi et al. “Simple data balancing achieves competitive worst-group-accuracy.” CLeaR 2022.