Efficient Bias Mitigation Without Privileged Information






Mateo Espinosa Zarlenga

Swami Sankaranarayanan

Jerone T. A. Andrews

Zohreh Shams

Mateja Jamnik

Alice Xiang

Sony AI

ECCV

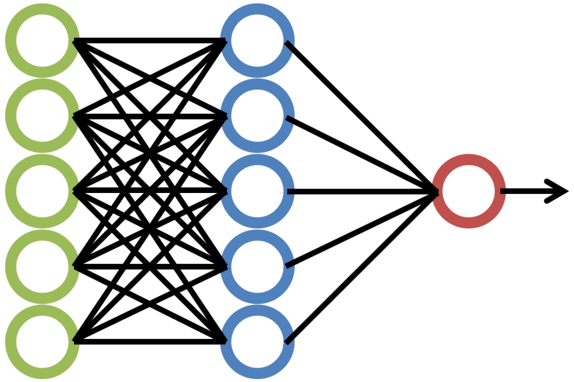
University of Cambridge

Supervised Learning's Bread And Butter

Today we will consider the traditional supervised learning setup:

Training Samples

$$\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathbb{R}^m, y_i \in \{1, 2, \cdots, L\}\}_{i=1}^{n}$$



Deep Neural Network (DNN)

$$f_\theta : \mathbb{R}^m \to [0, 1]^L$$

Supervised Learning's Bread And Butter

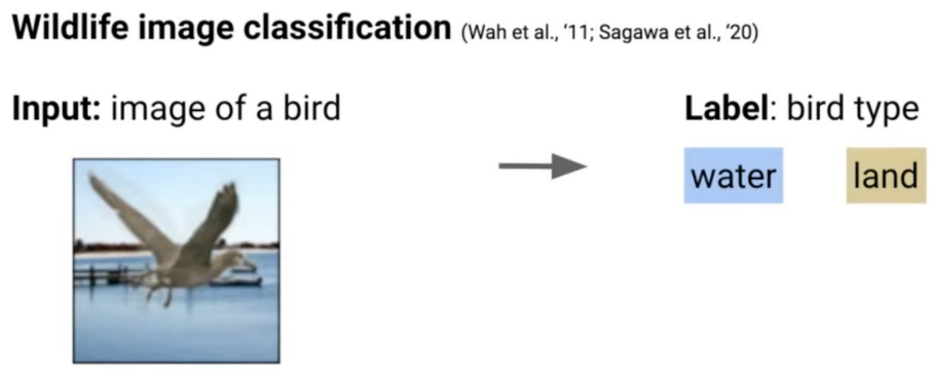
In particular, we will focus on instances when a DNN f_θ is trained by minimizing the empirical mean loss $\ell(\cdot)$ over the training set:

$$J_{ERM}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

This is what is referred to as Empirical Risk Minimization (ERM).

The Risks of Empirical Risk Minimization

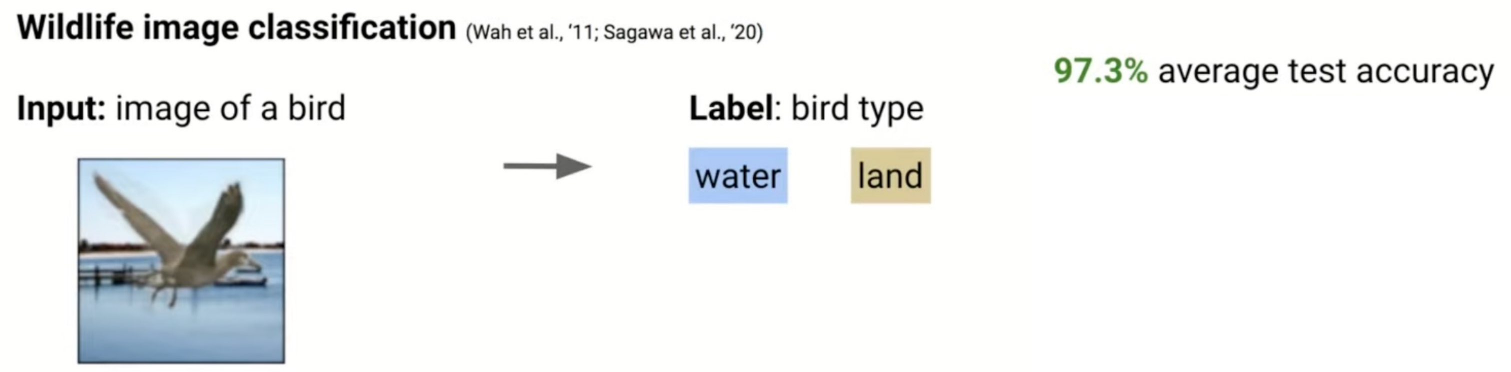
Even when ERM leads to high performance on average, this can change when we look at specific groups:



Example and slide adapted from [Liu et al., "Just Train Twice: Improving Group Robustness without Training Group Information" (2021)]

The Risks of Empirical Risk Minimization

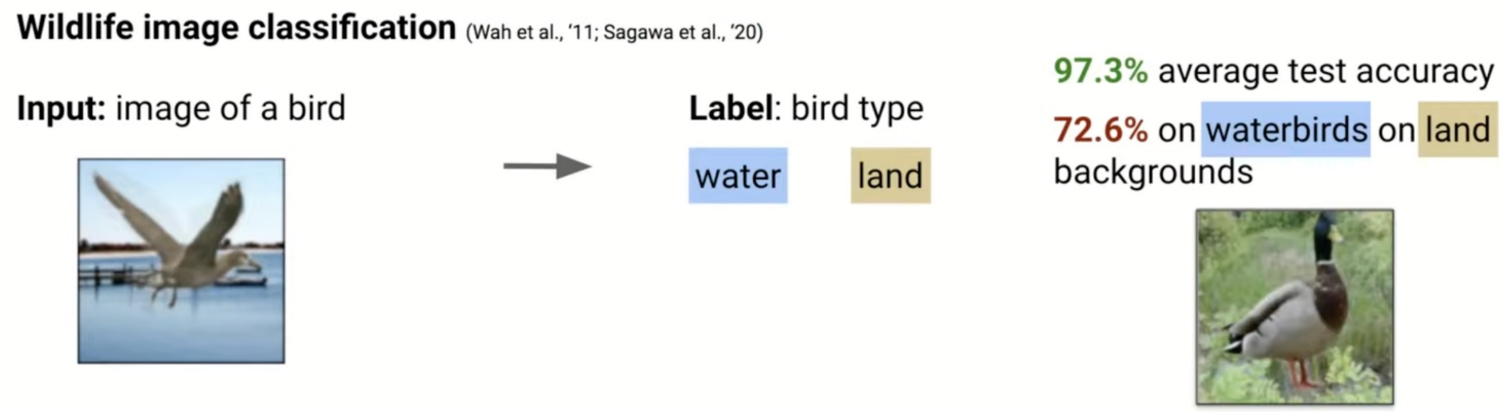
Even when ERM leads to high performance on average, this can change when we look at specific groups:



Example and slide adapted from [Liu et al., "Just Train Twice: Improving Group Robustness without Training Group Information" (2021)]

The Risks of Empirical Risk Minimization

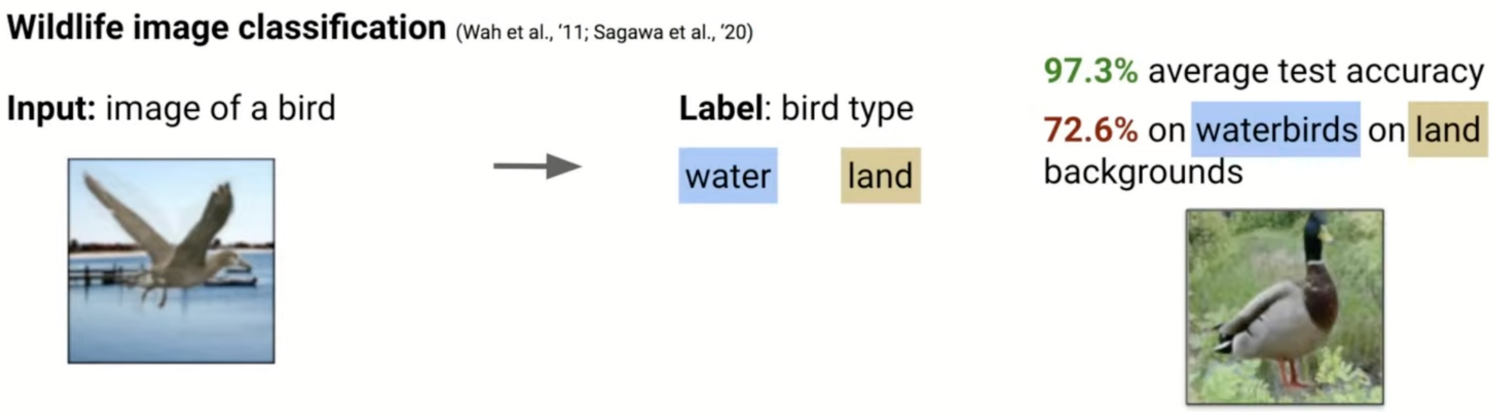
Even when ERM leads to high performance on average, this can change when we look at specific groups:



Example and slide adapted from [Liu et al., "Just Train Twice: Improving Group Robustness without Training Group Information" (2021)]

The Risks of Empirical Risk Minimization

Even when ERM leads to high performance on average, this can change when we look at specific groups:



We are therefore interested in maximizing the **Worst Group Accuracy (WGA)**:

$$\mathrm{WGA}(f_\theta, \mathcal{P}) := \min_{g \in \{1,2,\cdots,k\}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}(\mathbf{x},y|g)}\left[\mathbb{1}\left(f_\theta(\mathbf{x}) = y\right)\right]$$

Example and slide adapted from [Liu et al., "Just Train Twice: Improving Group Robustness without Training Group Information" (2021)]

Bias Mitigation

There is a plethora of bias mitigation methods

MaskTune: Mitigating Spurious Correlations by Forcing to Explore

Saeid Asgari Taghanaki* Autodesk AI Lab
Aliasghar Khani* Autodesk AI Lab
Fereshte Khani* Stanford University
Ali Gholami* Autodesk AI Lab
Linh Tran Autodesk AI Lab
Ali Mahdavi-Amiri Simon Fraser University
Ghassan Hamarneh Simon Fraser University

No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems

Nimit S. Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, Christopher Ré
Stanford University
nims@stanford.edu, jdunnmon@cs.stanford.edu, gdlangus@cs.stanford.edu, albertgu@stanford.edu, chrismre@cs.stanford.edu

Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization

Shiori Sagawa* Stanford University ssagawa@cs.stanford.edu
Pang Wei Koh* Stanford University pangwei@cs.stanford.edu
Tatsunori B. Hashimoto Microsoft tahashim@microsoft.com
Percy Liang Stanford University pliang@cs.stanford.edu

Multiaccuracy: Black-Box Post-Processing for Fairness in Classification

Michael P. Kim*† mpk@cs.stanford.edu Stanford University
Amirata Ghorbani† amiratag@stanford.edu Stanford University
James Zou jamesz@stan Stanford Un

Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations

Polina Kirichenko* New York University
Pavel Izmailov* New York University
Andrew Gordon Wilson New York University

Towards Last-layer Retraining for Group Robustness with Fewer Annotations

Tyler LaBonte[1] **Vidya Muthukumar**[2,1] **Abhishek Kumar**[3]
[1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology
[3]Google DeepMind
{tlabonte, vmuthukumar8}@gatech.edu abhishk@google.com

Learning from Failure: Training Debiased Classifier from Biased Classifier

Junhyun Nam[1] **Hyuntak Cha**[2] **Sungsoo Ahn**[1] **Jaeho Lee**[1] **Jinwoo Shin**[1,2]
[1]School of Electrical Engineering, KAIST
[2]Graduate School of AI, KAIST
{junhyun.nam, hyuntak.cha, sungsoo.ahn, jaeho-lee, jinwoos}@kaist.ac.kr

Just Train Twice: Improving Group Robustness without Training Group Information

Evan Zheran Liu *[1] **Behzad Haghgoo** *[1] **Annie S. Chen** *[1] **Aditi Raghunathan**[1] **Pang Wei Koh**[1] **Shiori Sagawa**[1] **Percy Liang**[1] **Chelsea Finn**[1]

Bias Mitigation

There is a plethora of bias mitigation methods

These can be:

1. **Group supervised**: we assume group labels during training

Bias Mitigation

There is a plethora of bias mitigation methods

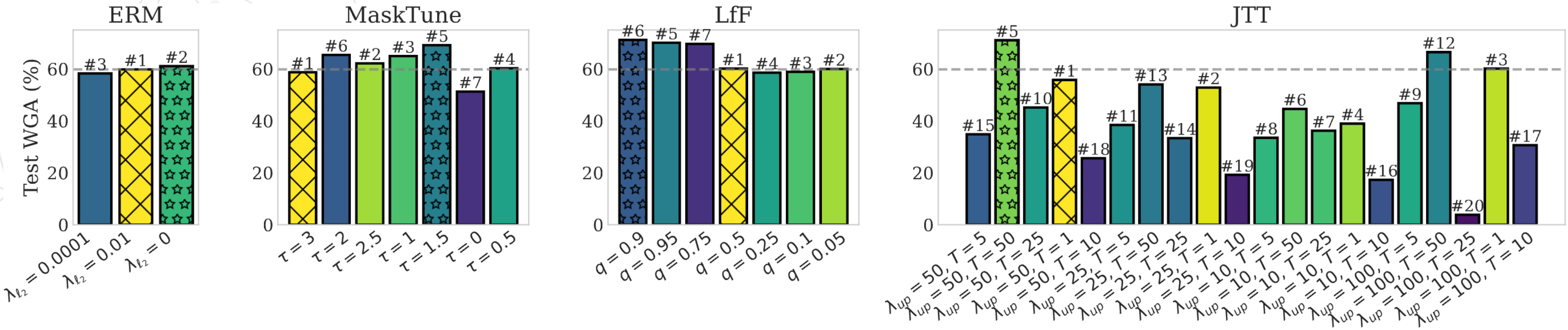
These can be:

1. **Group supervised**: we assume group labels during training

2. **Group Unsupervised**: we do not assume group labels during training*

*The Reality of Unsupervised Methods

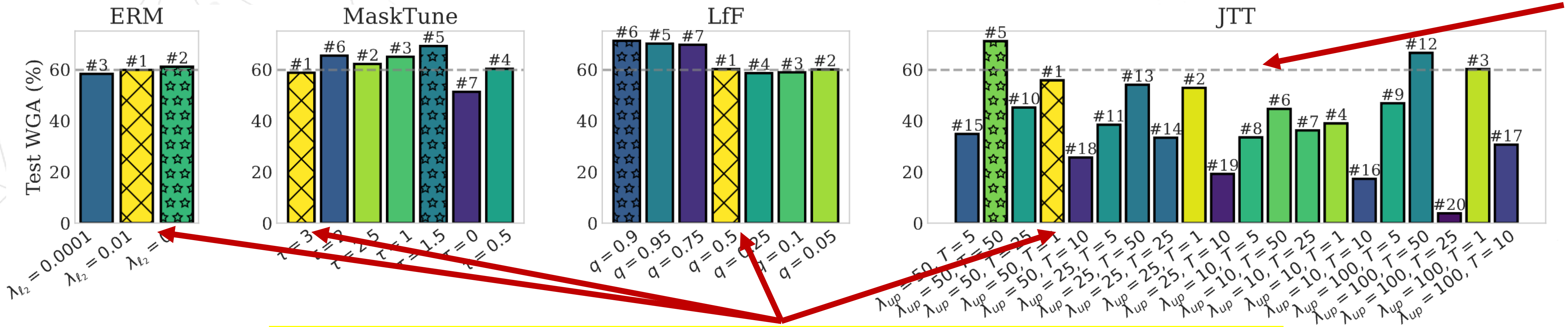
In practice, **unsupervised bias mitigation methods** need group labels during model selection to avoid selecting a biased model:



*The Reality of Unsupervised Methods

In practice, **unsupervised bias mitigation methods** need group labels during model selection to avoid selecting a biased model:

ERM Baseline



Selected model if validation group labels are unavailable

The selected hyperparameters lead to a model **no better than an ERM model**!

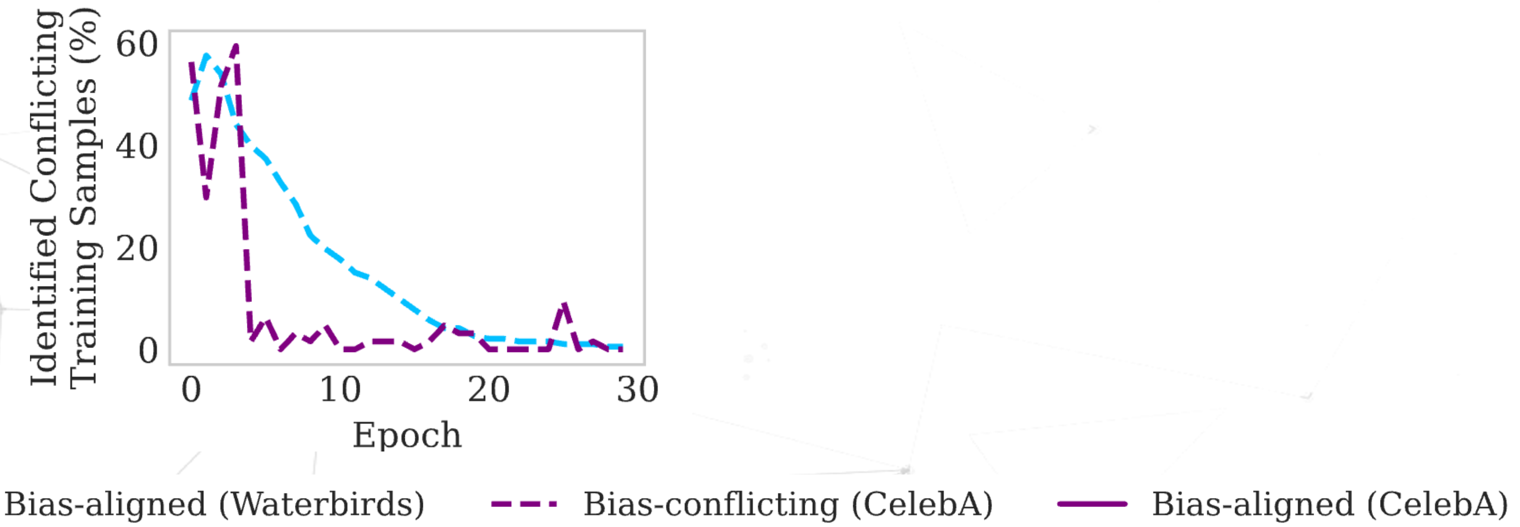
Our Work

How can we design a bias mitigation method that does not require group labels for either training or model selection?

Circumventing Privileged Information: Insights

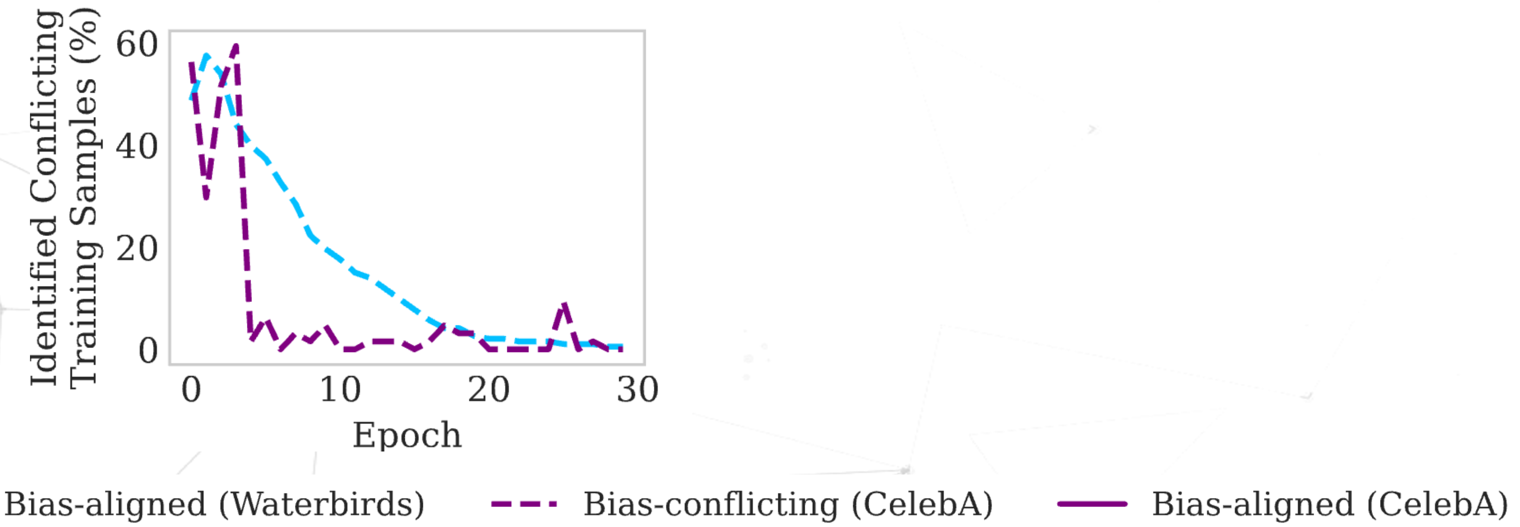
Circumventing Privileged Information: Insights

Insight #1 (Nam et al. and Liu et al.): Samples with spurious correlations are learnt before samples without the spurious correlation



Circumventing Privileged Information: Insights

Insight #1 (Nam et al. and Liu et al.): Samples with spurious correlations are learnt before samples without the spurious correlation

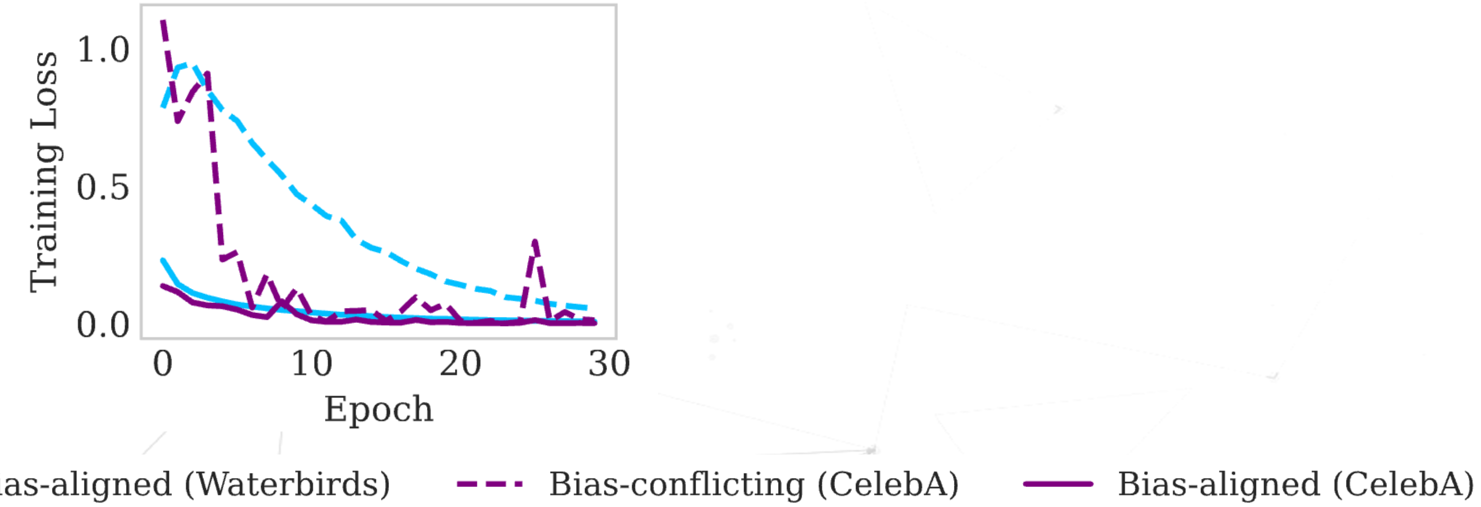


Previous works [1, 2] exploit this by looking at **a pre-determined "time slice"**

[1] Nam et al. "Learning from failure: De-biasing classifier from biased classifier." *NeurIPS 2020.*
[2] Liu et al. "Just train twice: Improving group robustness without training group information." *ICML*, 2021.

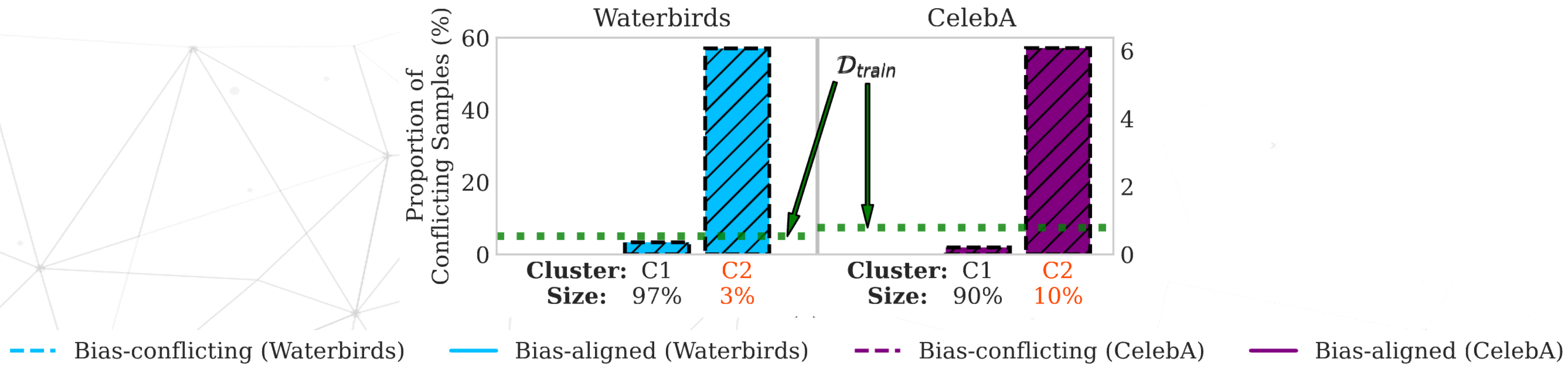
Circumventing Privileged Information: Insights

Insight #2: Training loss histories are very informative signals



Circumventing Privileged Information: Insights

Insight #2: Training loss histories are very informative signals

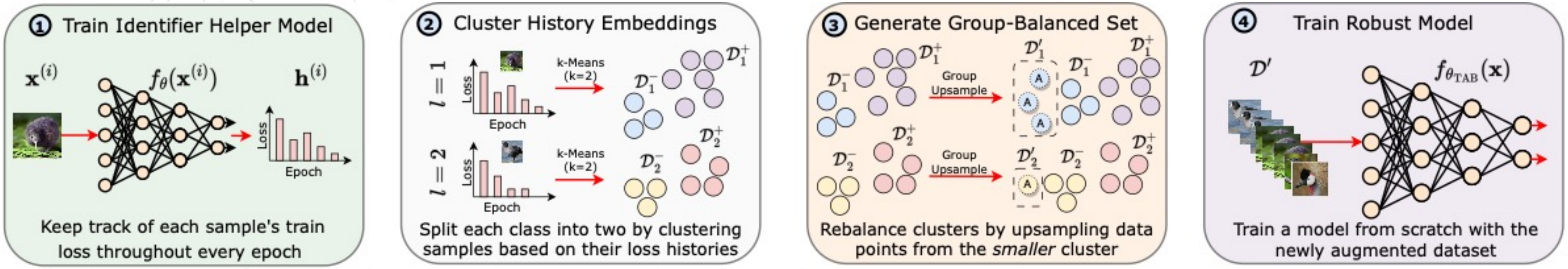


Clustering samples based on their training histories produces **a data subset with a higher proportion of samples without the spurious correlation!**

Targeted Augmentations for Bias Mitigation

Targeted Augmentations for Bias Mitigation

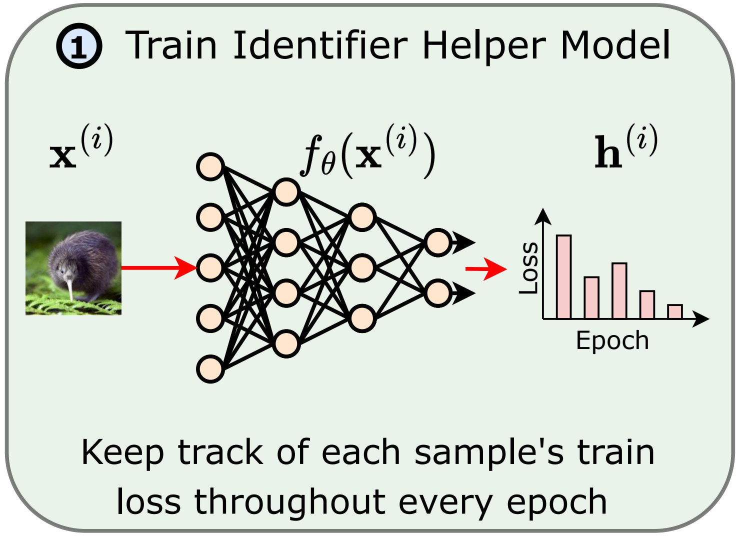
We propose **Targeted Augmentations for Bias mitigation (TAB)**, a new **hyperparameter-free** group-unsupervised bias mitigation pipeline:



Our approach exploits the **training history** of an identification model to generate **a group-balanced dataset** from which a robust model can be trained

Targeted Augmentations for Bias Mitigation

TAB first trains an ERM model while **keeping track of the loss** across all training samples and epochs:

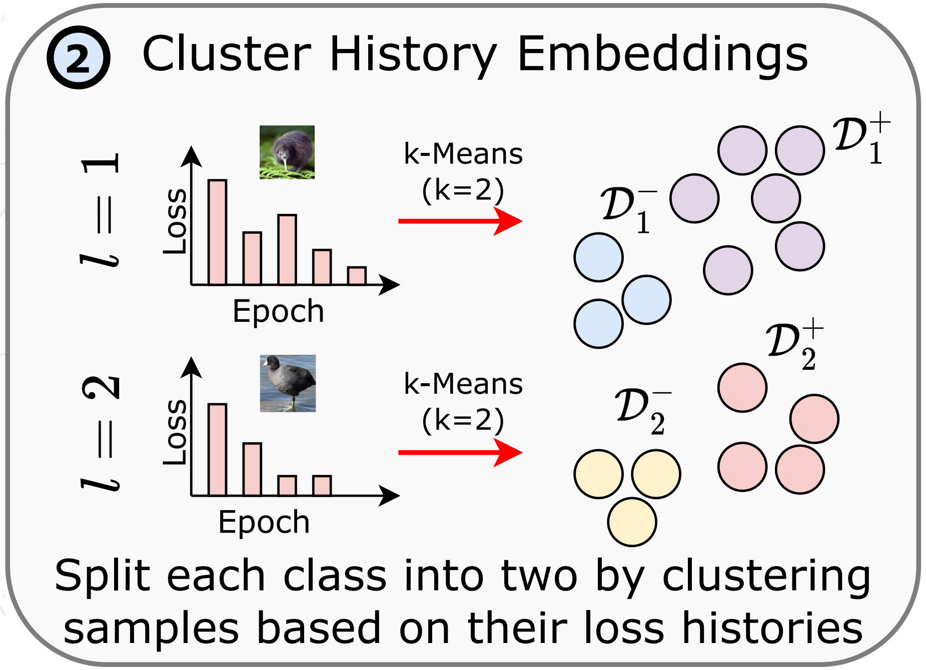


$$\boldsymbol{h}_i := \left[\ell\left(f_{\theta^{(1)}}(\boldsymbol{x}_i), y_i\right), \ell\left(f_{\theta^{(2)}}(\boldsymbol{x}_i), y_i\right), \cdots, \ell\left(f_{\theta^{(T)}}(\boldsymbol{x}_i), y_i\right)\right]^T$$

The loss during training provides valuable insights into which concepts are present or missing in a sample.

Targeted Augmentations for Bias Mitigation

We then identify error groups by **clustering the loss history embedding** space for each class label:

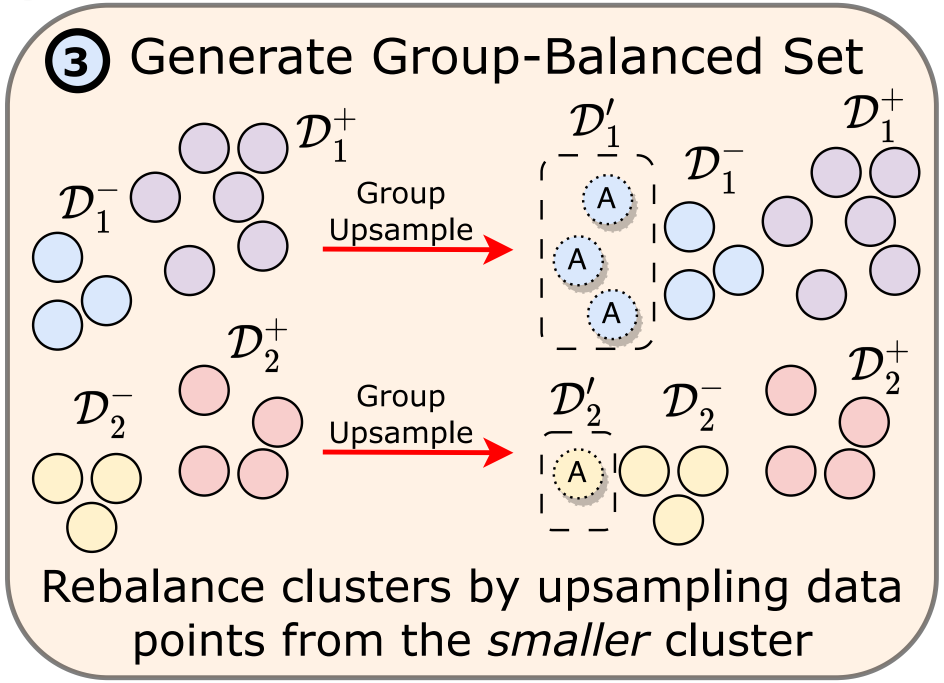


$g_i \coloneqq ClusterLabel(\boldsymbol{h}_i, 2\text{–}Means(H_l))$ for all $\mathbf{h_i} \in H_l$

where $H_l := \{h_j \mid y_j = l\}$ is the set of history embeddings for samples in class l

Targeted Augmentations for Bias Mitigation

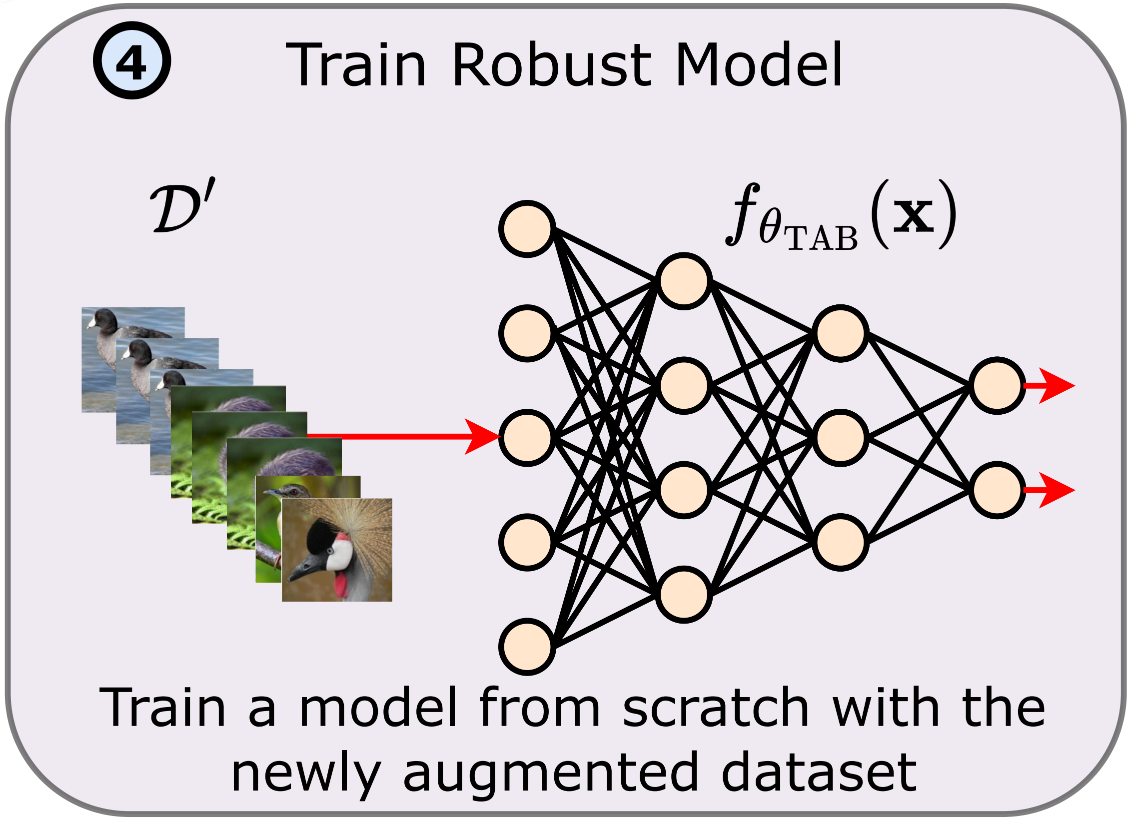
Next, we generate a **group-balanced training set** by **upsampling** each minority cluster to match the size of the majority cluster.



We do so by **randomly upsampling elements from the minority cluster** .

Targeted Augmentations for Bias Mitigation

Finally, we **train a robust model** using ERM on this group-balanced dataset:



Targeted Augmentations for Bias Mitigation

So how does TAB perform in practice?

Key Results TL;DR

	Method - {Hypers}	Even-Odd ($p = 99\%$)	cMNIST ($p = 98\%$)	Waterbirds	CelebA	BAR	CUB
WGA (%)	G-DRO - $\{\eta, \lambda_{\ell_2}\}$	57.66 ± 6.76	59.29 ± 3.27	68.54 ± 1.75	85.74 ± 0.69	N/A	N/A
	ERM - $\{\eta, \lambda_{\ell_2}\}$	55.98 ± 13.85	46.97 ± 8.71	44.86 ± 1.11	34.81 ± 0.26	29.56 ± 1.78	16.67 ± 0.00
	LfF - $\{q\}$	2.97 ± 3.36	48.45 ± 5.83	51.14 ± 1.08	$\mathbf{40.00 \pm 0.00}$	29.56 ± 2.35	14.44 ± 3.14
	JTT - $\{T, \lambda_{\text{up}}\}$	79.32 ± 1.76	57.21 ± 3.59	44.50 ± 0.45	37.78 ± 2.83	30.98 ± 2.00	12.22 ± 1.57
	MaskTune - $\{\tau\}$	72.82 ± 3.08	13.94 ± 7.37	35.67 ± 1.75	37.04 ± 1.14	17.61 ± 1.54	10.00 ± 7.20
	TAB (ours) - $\emptyset$	$\mathbf{81.85 \pm 2.39}$	$\mathbf{63.26 \pm 2.50}$	$\mathbf{55.92 \pm 1.80}$	$\mathbf{40.00 \pm 1.20}$	$\mathbf{38.94 \pm 1.03}$	$\mathbf{18.89 \pm 1.57}$
Mean Acc. (%)	G-DRO - $\{\eta, \lambda_{\ell_2}\}$	58.97 ± 6.79	94.83 ± 0.55	97.19 ± 0.28	92.67 ± 0.14	N/A	N/A
	ERM - $\{\eta, \lambda_{\ell_2}\}$	85.52 ± 12.09	91.22 ± 0.26	97.68 ± 0.06	$\mathbf{95.45 \pm 0.04}$	56.93 ± 1.13	$\mathbf{74.81 \pm 0.29}$
	LfF - $\{q\}$	60.29 ± 13.53	90.48 ± 1.17	97.46 ± 0.12	95.22 ± 0.02	55.96 ± 1.25	74.00 ± 0.67
	JTT - $\{T, \lambda_{\text{up}}\}$	93.12 ± 4.74	92.13 ± 1.13	97.71 ± 0.11	94.77 ± 0.05	58.00 ± 2.34	69.92 ± 0.10
	MaskTune - $\{\tau\}$	92.60 ± 5.02	83.25 ± 3.26	$\mathbf{98.15 \pm 0.04}$	95.32 ± 0.07	50.66 ± 1.38	70.07 ± 0.97
	TAB (ours) - $\emptyset$	$\mathbf{94.98 \pm 3.37}$	$\mathbf{93.28 \pm 1.09}$	97.52 ± 0.09	94.67 ± 0.05	$\mathbf{61.11 \pm 0.94}$	72.98 ± 0.34

TAB achieves better worst-group accuracies than competing approaches while maintaining a competitive mean accuracy compared to ERM models

Paper, poster, and contact information

Paper

Poster Information

Today, Tuesday Oct 1st

4:30 p.m. — 6:30 p.m.

Poster #27

Project Page

If you want to discuss further, contact me at me466@cam.ac.uk