

UNIVERSITY OF CAMBRIDGE

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

Polina Kirichenko, Pavel Izmailov & Andrew Gordon Wilson
ICLR 2023

Maximilian von Klinski

Machine Learning for Language Processing
Michaelmas 2026

The paper in one sentence

A network trained with plain ERM still learns the core features, even when it relies on spurious ones. Retraining only its last layer on a small group-balanced set restores worst-group accuracy.

- Deep Feature Reweighting (DFR): freeze the encoder, refit a linear head on held-out data where the shortcut does not predict the label
- Matches Group DRO-level worst-group accuracy on four benchmarks (two vision, two NLP) at a fraction of the cost
- **The problem sits in the last layer, not in the features**: robustness to a known shortcut becomes a linear problem

What is a spurious correlation?

PREMISE
A man is playing guitar on a stage.

HYPOTHESIS
Nobody is playing music.

Gold label: **contradiction**

Typical example

PREMISE
Two children are running in a park.

HYPOTHESIS
Kids are outdoors.

Gold label: **entailment**

Typical example

PREMISE
The shop was closed, so we walked on.

HYPOTHESIS
They did **not** go into the shop.

Gold label: **entailment**

A shortcut model says: contradiction

Minority group

- **Spurious feature:** predicts the label in the training data without being what the label depends on (negation → contradiction; water background → waterbird)
- **Group** = label × spurious attribute. The shortcut works in the majority groups and fails in the minority groups

Examples in the style of MultiNLI; negation shortcut: Gururangan et al. (2018), Sagawa et al. (2020)

The setup: four benchmarks, one metric

Dataset	Task	Spurious attribute	Model	ERM mean	ERM worst group
Waterbirds image	landbird vs waterbird	background	ResNet-50	98.1	**74.9**
CelebA image	blond vs not blond	gender	ResNet-50	95.3	**46.9**
MultiNLI text	entailment / neutral / contradiction	negation words	BERT-base	82.8	**65.9**
CivilComments text	toxic vs not toxic	identity mentions	BERT-base	92.1	**55.6**

1

Test accuracy (%). Kirichenko et al. (2023), Table 2

- **Worst-group accuracy** = lowest test accuracy over all groups (label × spurious attribute); the paper's headline metric
- ① ERM looks fine on average, but loses 17 to 48 points on its worst group
- Standard interpretation before this paper: the encoder learned the shortcut instead of the task, so retrain the whole model with a group-aware objective (Group DRO, JTT)

→ **Is that diagnosis right?**

Are the features broken, or the classifier?

Core feature = what the label depends on (the bird). **Spurious feature** = a correlated shortcut (the background)

What they did

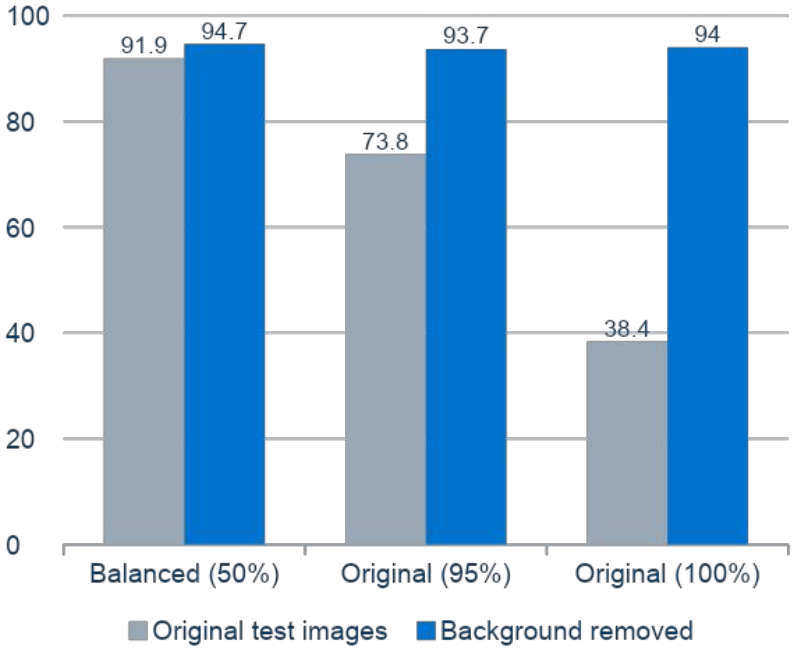
- Train a ResNet-50 with ERM on Waterbirds; background matches the label in 95% or 100% of images
- Test on the same images with the background removed

What they found

- Original test images: worst-group accuracy 74% and 38%
- Background removed: 94% in both cases, as good as a model trained on balanced data

What it means

- **The bird features are there;**
- **(the last layer) over-weights the background**
- **So fix the last layer, not the encoder**



Worst-group accuracy (%) by training set, Waterbirds.
Data: Kirichenko et al. (2023), Table 1

How does Deep Feature Reweighting work?

Two options

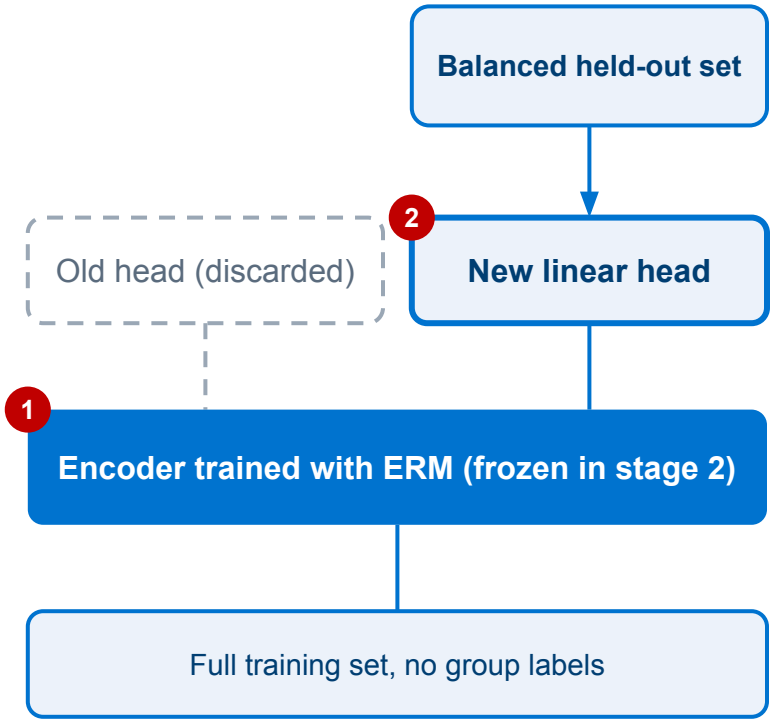
- **Robust training:** group labels for all training data, whole network retrained (Group DRO, Sagawa et al. 2020)
- **DFR:** keep the ERM model, refit only the linear head → this paper

The method in this paper

- ① Train the full model with ERM; no group labels, no reweighting
- ② Freeze the encoder; fit a new logistic-regression head (ℓ1) on a group-balanced subset of held-out data
- Average the head over 10 random balanced subsets; one hyperparameter (ℓ1 strength)

The aim

- **Down-weight features that only work when the shortcut is informative**



Example: Waterbirds has 2,048 ResNet-50 features and about 530 balanced validation images (133 per group)

Does it work? Benchmark results

Method	Group labels used	Waterbirds	CelebA	MultiNLI	CivilComm.
ERM	none	74.9	46.9	65.9	55.6
JTT	held-out: tuning only	86.7	81.1	72.6	69.3
SSA	held-out: training + tuning	89.0	**89.8**	76.6	69.9
Group DRO	all training data + tuning	91.4	88.9	**77.7** ②	69.9
DFR	held-out: training + tuning ①	**92.9**	88.3	74.7	**70.1**

Worst-group test accuracy (%), best per column in bold. Kirichenko et al. (2023), Table 2

What they found

- ① DFR gains 9 to 41 points over ERM; best on Waterbirds and CivilComments
- ② On MultiNLI it is 3 points behind Group DRO, 2 behind SSA
- Mean accuracy falls slightly (CivilComments 92.1 → 87.2)

What it means

- **A linear head on ERM features competes with full robust training**
- **"Matches", not "beats": the top methods are within 1 to 3 points**

What makes it work, and does it scale?

What they did

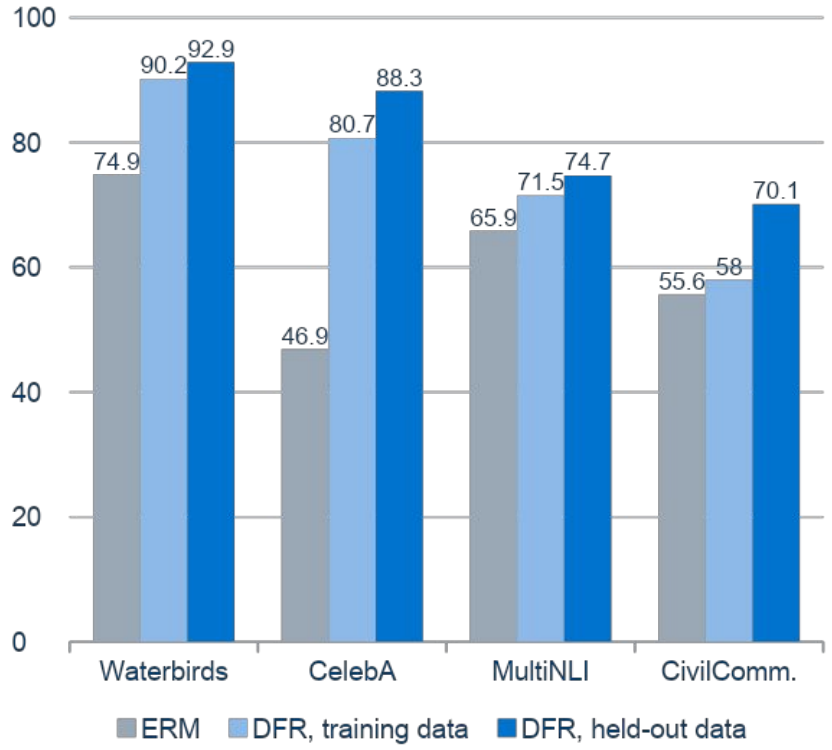
- Vary where the reweighting data come from and which layers are retrained
- Apply DFR to ImageNet models (background and texture reliance)

What they found

- Held-out data matter: DFR on training data is 3 to 12 points worse (chart)
- More layers do not help
- without ImageNet pre-training, Waterbirds drops to 54% (Tables 6, 7)
- ImageNet shape bias: 21% → 34%; full retraining reaches 81% (Table 3)

What it means

- **It needs pre-trained features and reweighting data the encoder has not seen**



Worst-group test accuracy (%).
Data: Kirichenko et al. (2023), Tables 2 and 9

What makes it work, and does it scale?

What they did

- Vary where the reweighting data come from and which layers are retrained
- Apply DFR to ImageNet models (background and texture reliance)

What they found

- Held-out data matter: DFR on training data is 3 to 12 points worse (chart)
- More layers do not help (Tables 7)
- without ImageNet pre-training, Waterbirds drops to 54% (Tables 6)
- ImageNet shape bias: 21% → 34%; full retraining reaches 81% (Table 3)

What it means

- **It needs pre-trained features and reweighting data the encoder has not seen**

Finetuned	WB	CelebA
Last Layer	$93.1_{\pm 0.2}$	$88.3_{\pm 0.5}$
Last 2 Layers	$93.1_{\pm 0.5}$	$87.2_{\pm 0.9}$
Last Block	$90.7_{\pm 0.4}$	$83.0_{\pm 0.9}$
All Layers	$35.6_{\pm 0.1}$	$77.8_{\pm 0.1}$

Worst-group test accuracy (%).
Data: Kirichenko et al. (2023), Tables 2 and 9

What makes it work, and does it scale?

What they did

- Vary where the reweighting data come from and which layers are retrained
- Apply DFR to ImageNet models (background and texture reliance)

What they found

- Held-out data matter: DFR on training data is 3 to 12 points worse (chart)
- More layers do not help (Tables 7)
- without ImageNet pre-training, Waterbirds drops to 54% (Tables 6)
- ImageNet shape bias: 21% → 34%; full retraining reaches 81% (Table 3)

What it means

- **It needs pre-trained features and reweighting data the encoder has not seen**

Method	ImageNet Pretrain	Dataset Fine-tune	Waterbirds		CelebA	
			Worst(%)	Mean(%)	Worst(%)	Mean(%)
Base Model	✓	✓	$74.9_{\pm 2.4}$	$98.1_{\pm 0.1}$	$46.9_{\pm 2.8}$	$95.3_{\pm 0}$
$\text{DFR}^{\text{Tr}}_{\text{Tr}}$	✓	✓	$90.2_{\pm 0.8}$	$97.0_{\pm 0.3}$	$80.7_{\pm 2.4}$	$85.4_{\pm 0.4}$
$\text{DFR}^{\text{Val}}_{\text{Tr}}$	✓	✓	$92.9_{\pm 0.2}$	$94.2_{\pm 0.4}$	$88.3_{\pm 1.1}$	$89.6_{\pm 0.4}$
Base Model	✗	✓	$6.9_{\pm 3.0}$	$88.0_{\pm 1.1}$	$39.8_{\pm 2.0}$	$95.7_{\pm 0.1}$
$\text{DFR}^{\text{Tr}}_{\text{Tr}}$	✗	✓	$45.4_{\pm 4.1}$	$69.8_{\pm 7.0}$	$83.4_{\pm 2.6}$	$87.5_{\pm 0.3}$
$\text{DFR}^{\text{Val}}_{\text{Tr}}$	✗	✓	$53.9_{\pm 1.8}$	$62.6_{\pm 2.2}$	$85.0_{\pm 2.1}$	$87.6_{\pm 0.3}$
$\text{DFR}^{\text{Val}}_{\text{IN}}$	✓	✗	$88.7_{\pm 0.4}$	$89.7_{\pm 0.1}$	$73.1_{\pm 2.6}$	$80.9_{\pm 0.5}$

Worst-group test accuracy (%).
Data: Kirichenko et al. (2023), Tables 2 and 9

What makes it work, and does it scale?

What they did

- Vary where the reweighting data come from and which layers are retrained
- Apply DFR to ImageNet models (background and texture reliance)

What they found

- Held-out data matter: DFR on training data is 3 to 12 points worse (chart)
- More layers do not help (Tables 7)
- without ImageNet pre-training, Waterbirds drops to 54% (Tables 6)
- ImageNet shape bias: 21% → 34%; full retraining reaches 81% (Table 3)

Method	Training Data	Shape bias (%)	Top-1 Acc (%)		
			ImageNet	ImageNet-R	ImageNet-C
RN-50	IN	21.4	76.0	23.8	39.8
	SIN	**81.4**	60.3	26.9	38.1
	IN+SIN	34.7	74.6	**27.6**	**45.7**
Shape-RN-50	IN+SIN	20.5	**76.8**	25.6	42.3
DFR	SIN	34.0	65.1	24.6	36.7
	IN+SIN	30.6	74.5	**27.2**	40.7

Worst-group test accuracy (%).
Data: Kirichenko et al. (2023), Tables 2 and 9

Critique: were the stated aims achieved?

Claim	Verdict	But…
ERM models learn core features even when they rely on spurious ones	**Yes, for vision**	• Feature analysis on images only; none for BERT • Largely inherited from pre-training: 54% on Waterbirds without it
Last-layer retraining matches state-of-the-art robustness methods	**Partly**	• Only when trained on held-out group labels; with Group DRO's data it is worse on all four tasks (CelebA 80.7 vs 88.9) • Behind Group DRO and SSA on MultiNLI
It reduces background and texture reliance on ImageNet	**Partly**	• Shape bias 34% vs 81% for full retraining; ImageNet-R barely moves

Technical reproducibility

Largely yes: public code (BSD), released checkpoints, full hyperparameters, 5 seeds, cheap to run; later papers reproduce the vision results within 1 to 2 points. **But:** the official repository has no MultiNLI or CivilComments code, there are no significance tests, and DFR results on CivilComments range from 70 to 82 across papers.

Limitations: stated and unstated

Stated by the authors

Needs a group-balanced held-out set with group labels. Falls well short of full retraining on Stylized ImageNet; little gain on ImageNet-R.

Not stated, but they matter for the claim

1. **Unequal data budget**
   DFR trains on held-out group labels that baselines only tune on. Waterbirds: 599 minority-group images in validation, 240 in the training set.

2. **The shortcut must be known**
   One annotated attribute per dataset. Real corpora hide several: negation is one NLI artefact among many (hypothesis-only cues, lexical overlap).

3. **"Core features are learned" may be pre-training**
   Evidence is vision-only. A pre-trained ConvNeXt reaches 94% on Waterbirds with no Waterbirds training at all (Izmailov et al. 2022).

4. **One metric, known groups only**
   Mean accuracy falls (CivilComments 92.1 → 87.2). No test of the BERT models on a new shift such as HANS; no significance tests for 1 to 3 point gaps.

What later work found

1 Why it works is disputed

Hill et al. (2025): the gain tracks the group balance of the held-out set; with the training group ratio, last-layer retraining does not beat ERM

LaBonte et al. (2023): class balancing alone gives 94% of DFR's worst-group accuracy

2 Not always sufficient

Le et al. (2023): on skin-lesion images DFR still attends to the spurious patch

Yang et al. (2023): last-layer methods help with spurious correlation, not with attribute imbalance or unseen attributes

Chew et al. (2023), text: fine-tuning moves spurious tokens next to label words in embedding space

3 The benchmarks are shaky

Bell et al. (2024): benchmarks disagree; DFR is best on Waterbirds, second-worst on NICO++; MultiNLI fails a validity check

Izmailov et al. (2022): pre-trained features alone nearly solve Waterbirds

Common thread: DFR does raise worst-group accuracy when a balanced, annotated held-out set exists. The dispute is over “sufficient”, and over what these benchmarks measure.

How does it advance the state of the art?

- **It relocates the problem: from feature learning to the last layer**
  - a cheap, strong baseline that new robustness methods must beat
  - DFR doubles as a probe of how much core information features hold
- **It questions robust-training practice**
  - Group DRO and JTT mostly fix the last layer, not the features (Izmailov et al. 2022)
  - group labels may be better spent on a small held-out set
- **Questions taken up by later work**
  - Qiu et al. 2023 (AFR): the same idea without group labels
  - LaBonte et al. 2023 (SELF): a few annotations suffice
  - Hill et al. 2025: is it only the group balance?

Top 25%

notable paper at ICLR 2023

640+

citations (alphaXiv count, November 2026)

References

- Kirichenko, Izmailov & Wilson (2023). Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. ICLR. arXiv:2204.02937. Code: github.com/PolinaKirichenko/deep_feature_reweighting
- Sagawa, Koh, Hashimoto & Liang (2020). Distributionally Robust Neural Networks for Group Shifts. ICLR.
- Izmailov, Kirichenko, Gruver & Wilson (2022). On Feature Learning in the Presence of Spurious Correlations. NeurIPS. arXiv:2210.11369
- Qiu, Potapczynski, Izmailov & Wilson (2023). Simple and Fast Group Robustness by Automatic Feature Reweighting. ICML. arXiv:2306.11074
- LaBonte, Muthukumar & Kumar (2023). Towards Last-layer Retraining for Group Robustness with Fewer Annotations. NeurIPS. arXiv:2309.08534
- Yang, Zhang, Katabi & Ghassemi (2023). Change is Hard: A Closer Look at Subpopulation Shift. ICML. arXiv:2302.12254
- Le, Schlötterer & Seifert (2023). Is Last Layer Re-Training Truly Sufficient for Robustness to Spurious Correlations? arXiv:2308.00473
- Chew, Lin, Chang & Huang (2023). Understanding and Mitigating Spurious Correlations in Text Classification with Neighborhood Analysis. arXiv:2305.13654
- Bell, Bouchacourt & Sagun (2024). Reassessing the Validity of Spurious Correlations Benchmarks. arXiv:2409.04188
- Hill, LaBonte, Zhang & Muthukumar (2025). On the Unreasonable Effectiveness of Last-layer Retraining. arXiv:2512.01766
- Gururangan et al. (2018). Annotation Artifacts in Natural Language Inference Data. NAACL. McCoy, Pavlick & Linzen (2019). Right for the Wrong Reasons (HANS). ACL.

Further Information: do the numbers replicate?

DFR worst-group accuracy (%) reported by	Waterbirds	CelebA	MultiNLI	CivilComm.
Original paper, Table 2	92.9	88.3	74.7	70.1
Authors' released checkpoints (code README)	92.0	88.0	–	–
LaBonte et al. 2023, values cited for DFR	91.1	89.4	72.6	78.8
LaBonte et al. 2023, own implementation	92.4	87.0	70.8	81.8
Le et al. 2023, reproduction	92.6	–	–	–
Yang et al. 2023, no group labels (other protocol)	–	–	63.8	64.4

1

- Vision results are stable across independent groups (within 1 to 2 points)
- MultiNLI comes out 1 to 4 points lower elsewhere
- ① CivilComments varies by up to 12 points across papers