# Improving Group Robustness on Spurious Correlation via Evidential Alignment (Extended Abstract)*

**Wenqian Ye**, **Guangtao Zheng**, **Aidong Zhang**
University of Virginia, Charlottesville, VA, USA
{wenqian, gz5hp, aidong}@virginia.edu

## Abstract

Deep neural networks often rely on spurious correlations, i.e., superficial associations between non-causal features and prediction targets. While yielding high overall accuracy during training, such reliance degrades generalization on minority groups where these correlations break down. Existing methods mitigate this issue by using external group annotations or auxiliary deterministic models, but group annotations are costly to obtain and deterministic auxiliaries may fail to capture the full spectrum of biases learned by the model. We propose Evidential Alignment, a framework that leverages uncertainty quantification to identify and suppress spurious correlations without requiring group annotations. By transforming a biased ERM model from first-order to second-order risk minimization with a Dirichlet distribution and applying an evidential calibration step that reweighs samples by their epistemic uncertainty, Evidential Alignment debiases the model while preserving core features. We provide theoretical guarantees and demonstrate strong worst-group accuracy across diverse architectures and modalities.

## 1 Introduction

Deep neural networks trained with empirical risk minimization (ERM) [Vapnik, 1999] often exhibit *spurious biases*, i.e., reliance on brittle, non-causal features such as background, texture, or secondary objects [Stock and Cisse, 2018; Beery *et al.*, 2018; Sagawa *et al.*, 2020; Ye *et al.*, 2024]. For example, in a cow/camel classification task, the camel class can be spuriously correlated with desert backgrounds, leading ERM-trained models to fail on minority groups (e.g., camels on grass) [Beery *et al.*, 2018]. Such failures are particularly problematic in high-stakes domains including medicine [Zech *et al.*, 2018], fairness [Khani and Liang, 2021], and criminal justice [Barocas *et al.*, 2023].

A model's robustness to such failures is typically measured by its *worst-group accuracy* (WGA) [Sagawa *et al.*, 2020]. Supervised approaches such as Group DRO [Sagawa *et al.*, 2020], DFR [Kirichenko *et al.*, 2023], and JTT [Liu *et al.*, 2021] require explicit group labels that pair class labels with spurious attributes, but such annotations are costly and often miss subtle biases. Recent annotation-free methods [Nam *et al.*, 2020; Li *et al.*, 2024] sidestep this requirement by using deterministic auxiliary models to amplify or pseudo-label biased samples. However, deterministic auxiliaries provide only a narrow view of how a model exploits spurious features, limiting their effectiveness.

We propose Evidential Alignment, a framework that leverages *epistemic uncertainty* to provide a principled, annotation-free signal for identifying and mitigating spurious biases. Drawing on the Dempster–Shafer Theory of Evidence [Yager and Liu, 2010] and recent advances in evidential deep learning [Sensoy *et al.*, 2018], we transform a biased ERM model from *first-order risk minimization* to *second-order risk minimization* by learning a distribution over distributions via a Dirichlet parameterization. The resulting epistemic uncertainty is then used to reweight calibration samples and retrain the classifier, suppressing spurious correlations while preserving core features.

The main **contributions** of this paper are:

- We propose Evidential Alignment, a novel framework that integrates second-order risk minimization with evidential calibration to leverage model uncertainty for identifying and mitigating spurious correlations *without* group annotations.
- We provide theoretical insights showing how uncertainty quantification captures spurious biases in latent space.
- Extensive experiments across diverse architectures and modalities demonstrate that Evidential Alignment significantly enhances group robustness and generalization.[1]

## 2 Related Work

**Group robustness.** Methods for improving group robustness against spurious correlations differ mainly in how much group supervision they assume. With full training-set group annotations, Group DRO [Sagawa *et al.*, 2020] minimizes

*Full version appears in the Proceedings of KDD 2025 [Ye *et al.*, 2025], recipient of the Best Paper Award in Research.

[1]Code: https://github.com/wenqian-ye/evidential_alignment.

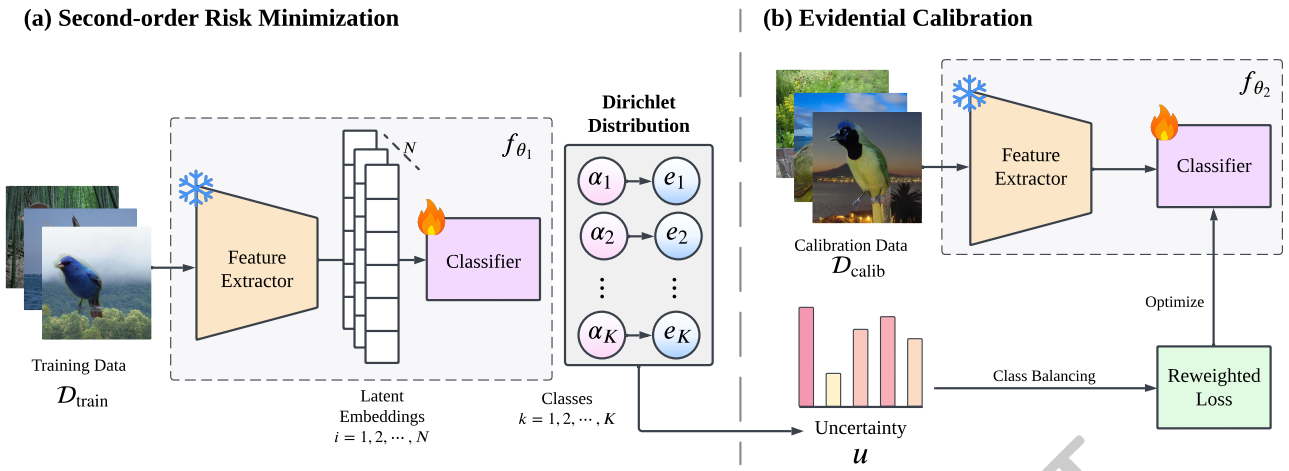


Figure 1: Overall framework of Evidential Alignment. **(a)** Train the classifier together with Dirichlet distribution parameters using second-order risk minimization on $\mathcal{D}_{\text{train}}$. **(b)** Compute the epistemic uncertainty for each calibration sample in $\mathcal{D}_{\text{calib}}$ and apply a reweighted, class-balanced loss to refine the classifier.

the worst-case loss across predefined groups. With only validation-set annotations, DFR [Kirichenko *et al.*, 2023] retrains the final layer on a group-balanced held-out set, and JTT [Liu *et al.*, 2021] as well as AFR [Qiu *et al.*, 2023] upweight samples that a biased ERM model misclassifies. Annotation-free approaches such as LfF [Nam *et al.*, 2020] and BAM [Li *et al.*, 2024] train a deliberately biased auxiliary model and use its errors as a proxy signal. SELF [LaBonte *et al.*, 2024] retrains the last layer with held-out data but without group labels. All of these annotation-free strategies rely on *deterministic* bias estimators, which capture only a narrow slice of how a model exploits spurious features.

**Uncertainty quantification.** Predictive uncertainty is typically decomposed into aleatoric (data) and epistemic (model) components [Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021]. Bayesian methods, MC dropout, and deep ensembles provide such estimates but require sampling or multiple model passes. Evidential deep learning [Sensoy *et al.*, 2018], grounded in the Dempster–Shafer theory of evidence [Yager and Liu, 2010], instead places a Dirichlet distribution over class probabilities in a single forward pass. Uncertainty has been used for calibration, active learning, and out-of-distribution detection, but its potential for diagnosing and correcting spurious bias has remained largely unexplored. Our method closes this gap.

# 3 Method

## 3.1 Problem Setup

We consider a classification problem with training set $\mathcal{D}_{\text{train}} = \{(x, y)\}$ partitioned into latent groups $\mathcal{D}_g^{\text{tr}}$, where each group $g := (y, a)$ pairs a class $y \in \mathcal{Y}$ with a spurious attribute $a \in \mathcal{A}$. Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be a model with parameters $\theta = (\phi_\theta, h_\theta)$ decomposed into a feature extractor $\phi_\theta$ and a classifier $h_\theta$. Standard ERM minimizes the average training loss; group robustness instead targets the worst-group risk

$$\min_{\theta \in \Theta} \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \mathcal{D}_g^{\text{tr}}} [\ell(f_\theta(x), y)]. \tag{1}$$

**Uncertainty decomposition.** The posterior predictive distribution can be split into aleatoric (data) and epistemic (model) parts by marginalizing over $\theta$:

$$p(y \mid x) = \int \underbrace{P(y \mid x, \theta)}_{\text{aleatoric}} \underbrace{p(\theta \mid \mathcal{D})}_{\text{epistemic}} \, d\theta. \tag{2}$$

For neural networks this integral is intractable. Following [Malinin and Gales, 2018; Ulmer *et al.*, 2021; Hüllermeier and Waegeman, 2021], we factorize it through an intermediate distribution $\boldsymbol{\pi}$ over class probabilities and use a point estimate $\hat{\theta}$, yielding $p(y \mid x) = \int P(y \mid \boldsymbol{\pi})\, p(\boldsymbol{\pi} \mid x, \hat{\theta})\, d\boldsymbol{\pi}$, which has a closed-form solution when $p(\boldsymbol{\pi} \mid x, \hat{\theta})$ is a Dirichlet distribution. We therefore focus on the *epistemic* component, which can be efficiently captured by a Dirichlet over class probabilities [Sensoy *et al.*, 2018; Hüllermeier and Waegeman, 2021] and provides the signal we use to debias $f_\theta$ without group labels.

## 3.2 Second-Order Risk Minimization

Conventional ERM optimizes a deterministic predictive distribution $p(y \mid \theta, x)$, conflating aleatoric and epistemic uncertainty. We instead model a **distribution over distributions** $p(\theta \mid \boldsymbol{\pi}(x))$, where $\boldsymbol{\pi}(x)$ parameterizes uncertainty about the prediction outcomes. We extend the ERM model to output non-negative evidence $e_k(x) \geq 0$ for each class $k \in \{1, \ldots, K\}$, defining the Dirichlet parameters

$$\alpha_k(x) = e_k(x) + 1, \quad k = 1, \ldots, K. \tag{3}$$

The Dirichlet distribution over class probabilities $\mathbf{p}(x)$ is

$$\text{Dir}(\mathbf{p}(x) \mid \boldsymbol{\alpha}(x)) = \frac{1}{B(\boldsymbol{\alpha}(x))} \prod_{k=1}^{K} p_k(x)^{\alpha_k(x)-1}, \tag{4}$$

with expected class probability $\mathbb{E}[p_k(x)] = \alpha_k(x)/S(x)$, where $S(x) = \sum_k \alpha_k(x)$.

The **second-order risk minimization** loss for a sample $(x, y)$ combines a classification objective with an evidence regularizer:

$$\mathcal{L}_1(x, y \mid \theta_1) = \underbrace{-\log \mathbb{E}[p_y(x)]}_{\text{classification}} + \lambda_t \underbrace{\mathrm{KL}\big(\mathrm{Dir}(\boldsymbol{\alpha}(x)) \,\|\, \mathrm{Dir}(\mathbf{1})\big)}_{\text{evidence regularization}}, \quad (5)$$

where $\mathrm{Dir}(\mathbf{1})$ is a non-informative uniform Dirichlet prior. The classification objective is the negative log expected probability of the true class,

$$\mathcal{L}_{\text{cls}}(x, y \mid \theta_1) = -\log \frac{\alpha_y(x)}{S(x)} = -\log \frac{e_y(x) + 1}{\sum_{k=1}^{K} \big(e_k(x) + 1\big)}, \quad (6)$$

and the KL divergence between the predicted Dirichlet and the uniform prior has the closed form

$$\mathrm{KL}\big(\mathrm{Dir}(\boldsymbol{\alpha}) \,\|\, \mathrm{Dir}(\mathbf{1})\big) = \log \frac{B(\mathbf{1})}{B(\boldsymbol{\alpha})} + \sum_{k=1}^{K} (\alpha_k - 1)\big(\psi(\alpha_k) - \psi(S)\big), \quad (7)$$

where $\psi(\cdot) = \frac{d}{dz} \ln \Gamma(\cdot)$ is the digamma function. The KL term penalizes overconfidence and promotes calibrated uncertainty. We adopt a dynamic annealing coefficient $\lambda_t = \min(t/\eta, 1)$ to stabilize training, where $t$ is the current epoch and $\eta$ is the annealing-step hyperparameter.

### 3.3 Evidential Calibration

After second-order training, the learned Dirichlet yields a per-sample epistemic uncertainty

$$u(x) = \frac{K}{S(x)} = \frac{K}{\sum_{k=1}^{K} \alpha_k(x)}. \quad (8)$$

On a held-out calibration set $\mathcal{D}_{\text{calib}}$, we upweight misclassified samples in proportion to their uncertainty,

$$w(x, y) = \mathbb{1}[f_{\theta_1}(x) = y] + u(x) \cdot \mathbb{1}[f_{\theta_1}(x) \neq y], \quad (9)$$

and retrain only the last layer with a class-balanced reweighted cross-entropy plus a proximal regularizer:

$$\mathcal{L}_2(\mathcal{D}_{\text{calib}} \mid u, \theta_2) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\text{calib}}} \big[w(x, y)\, \ell_{\text{ce}}(f_{\theta_2}(x), y)\big] + \beta \|\theta_2 - \theta_1\|_2^2. \quad (10)$$

Last-layer retraining suffices to learn invariant features, following [Kirichenko *et al.*, 2023; LaBonte *et al.*, 2024]. The two stages are illustrated in Figure 1.

### 3.4 Theoretical Analysis

We establish two guarantees; full statements and detailed proofs are in the original paper.

**Theorem 1** (ELBO for Second-Order Risk)**.** *Let $p(\boldsymbol{\pi} \mid y)$ be the true posterior over class probabilities given $y$, and let $p(\boldsymbol{\pi} \mid x, \theta)$ be the variational evidential distribution. Training with $\mathcal{L}_1$ optimizes the evidence lower bound (ELBO) on $\log p(y)$, so second-order risk minimization is equivalent to maximizing the log marginal likelihood of observed labels under a variational Dirichlet posterior.*

**Theorem 2** (PAC-Bayes Bound for Reweighted Empirical Risk)**.** *Let $\mathcal{H}$ have $[0, 1]$-bounded loss and $\mathcal{G}$ be a finite set of latent groups with $n_g$ i.i.d. samples each, $n_{\min} = \min_g n_g$. For any reweighting $\{w_g\}$ with $\alpha = \min_g w_g > 0$ and any prior $P$ and posterior $Q$ over $\mathcal{H}$, the worst-group risk satisfies, with probability at least $1 - \delta$,*

$$\max_g \mathbb{E}_{f\sim Q}[R_g(f)] \le \tfrac{1}{\alpha} \mathbb{E}_{f\sim Q}[\hat{R}_w(f)] + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln(|\mathcal{G}|/\delta)}{2n_{\min}}}.$$

Together, Theorem 1 certifies that our second-order objective is principled, since it maximizes a likelihood by treating evidence as the parameters of a variational Dirichlet posterior. Theorem 2 certifies that any reweighting whose induced empirical risk is small bounds the worst-group risk; our uncertainty-based weighting is exactly such a scheme, since high-uncertainty (likely-minority) samples are upweighted, raising $\alpha = \min_g w_g$ and tightening the bound.

## 4 Experiments

**Datasets.** We evaluate on four standard group-robustness benchmarks. **Waterbirds** [Sagawa *et al.*, 2020; Welinder *et al.*, 2010] contains waterbird/landbird images whose label is spuriously correlated with the water/land background; the minority groups are waterbirds on land and landbirds on water. **CelebA** [Liu *et al.*, 2015] requires predicting hair color (blond vs. non-blond), which is spuriously correlated with gender, with blond males being a tiny minority. **MultiNLI** [Williams *et al.*, 2017] is a sentence-pair entailment task whose contradiction label is spuriously correlated with the presence of negation words. **CivilComments** [Borkan *et al.*, 2019; Koh *et al.*, 2021] is a toxicity-classification task whose label spuriously correlates with eight demographic identity mentions; we use the WILDS split.

**Setup.** For images we use a ResNet-50 [He *et al.*, 2016] backbone pretrained on ImageNet; for text we use BERT-base [Kenton and Toutanova, 2019]. We form $\mathcal{D}_{\text{calib}}$ from one half of the validation split and reserve the other half for model selection; this matches the protocol of DFR$^\dagger$ and SELF$^\dagger$, but unlike DFR *we use no group labels at any stage*. Following [Kirichenko *et al.*, 2023], only the last layer is updated in both stages of our method. Baselines fall into two groups: methods that use validation-set group labels (JTT, AFR, DFR$^\dagger$, SELF$^\dagger$) and methods that use no group labels (LfF, BAM, Ours$^\dagger$). We report worst-group accuracy (WGA), average accuracy (Acc.), and their gap, averaged over three random seeds.

| Method | Group Annot. | | Waterbirds | | | CelebA | | |
|---|---|---|---|---|---|---|---|---|
| | Tr. | Val | WGA | Acc. | Gap | WGA | Acc. | Gap |
| ERM [Vapnik, 1999] | – | – | 72.6 | 97.3 | 24.7 | 47.2 | 95.6 | 48.4 |
| JTT [Liu *et al.*, 2021] | No | Yes | 86.7 | 93.3 | 6.6 | 81.1 | 88.0 | 6.9 |
| AFR [Qiu *et al.*, 2023] | No | Yes | 90.4 | 94.2 | 3.8 | 82.0 | 91.3 | 9.3 |
| DFR† [Kirichenko *et al.*, 2023] | No | Yes | 92.9 | 94.2 | 2.5 | 88.3 | 91.3 | 3.0 |
| SELF† [LaBonte *et al.*, 2024] | No | Yes | 93.0 | 94.0 | 1.0 | 83.9 | 91.7 | 7.8 |
| LfF [Nam *et al.*, 2020] | No | No | 78.0 | 91.2 | 13.2 | 77.2 | 85.1 | 7.9 |
| BAM [Li *et al.*, 2024] | No | No | 89.1 | 91.4 | 2.3 | 80.1 | 88.4 | 8.3 |
| **Ours**† | No | No | **92.2** | 95.1 | 2.9 | **84.4** | 92.3 | 7.9 |

Table 1: Worst-group accuracy (WGA), average accuracy (Acc.), and accuracy gap (Gap) across image datasets (ResNet-50). Best WGAs for annotation-free methods in bold. † uses a fraction of validation data for fine-tuning.

| Method | Group Annot. | | MultiNLI | | | CivilComments | | |
|---|---|---|---|---|---|---|---|---|
| | Tr. | Val | WGA | Acc. | Gap | WGA | Acc. | Gap |
| ERM [Vapnik, 1999] | – | – | 67.9 | 82.4 | 14.5 | 57.4 | 92.6 | 35.2 |
| JTT [Liu *et al.*, 2021] | No | Yes | 72.6 | 78.6 | 6.0 | 69.3 | 91.1 | 21.8 |
| AFR [Qiu *et al.*, 2023] | No | Yes | 73.4 | 81.4 | 8.0 | 68.7 | 89.8 | 21.1 |
| DFR† [Kirichenko *et al.*, 2023] | No | Yes | 70.8 | 81.7 | 10.9 | 81.8 | 87.5 | 5.7 |
| SELF† [LaBonte *et al.*, 2024] | No | Yes | 70.7 | 81.2 | 10.5 | 79.1 | 87.7 | 8.6 |
| LfF [Nam *et al.*, 2020] | No | No | 70.2 | 80.8 | 10.6 | 58.8 | 92.5 | 33.7 |
| BAM [Li *et al.*, 2024] | No | No | 70.8 | 80.3 | 9.5 | 79.3 | 88.3 | 9.0 |
| **Ours**† | No | No | **74.5** | 80.6 | 7.3 | **80.2** | 87.1 | 6.9 |

Table 2: Worst-group accuracy (WGA), average accuracy (Acc.), and accuracy gap (Gap) across text datasets (BERT-base). Best WGAs for annotation-free methods in bold. † uses a fraction of validation data for fine-tuning.

**Main results.** Tables 1 and 2 compare Evidential Alignment with state-of-the-art baselines. Among methods that use *no* group annotations (bottom blocks), our method achieves the best WGA on all four benchmarks, lifting Waterbirds from 72.6% (ERM) to 92.2% and CelebA from 47.2% to 84.4%. Concretely, on Waterbirds we improve over LfF by 14.2 points and BAM by 3.1 points; on CelebA, the gain over LfF is 7.2 points and over BAM 4.3 points. The trend persists in language: on MultiNLI we surpass the strongest annotation-free baseline (BAM) by 3.7 points, and on CivilComments we set a new annotation-free record at 80.2% WGA. We even match or exceed several methods that rely on group labels for validation: we outperform AFR, JTT, and SELF on most datasets, and remain competitive with DFR despite DFR's access to a group-balanced calibration set. Beyond accuracy, our method also yields a small accuracy gap (e.g. 2.9 on Waterbirds and 6.9 on CivilComments), indicating balanced performance across groups rather than skewed majority versus minority trade-offs. The full paper additionally reports a Colored MNIST synthetic study, results on the safety-critical CheXpert chest-radiograph dataset, and ViT-base experiments on Waterbirds, all confirming the same trend across architectures and modalities. Computationally, Evidential Alignment retrains only the last layer in both stages, so its overhead is on par with DFR and AFR rather than with JTT [Liu *et al.*, 2021], which retrains a full auxiliary model from scratch.

**Synthetic study.** On a Colored MNIST setup [Vapnik, 1999] where 90% of training samples carry a class–color spurious correlation but only 10% of test samples do, ERM collapses to 3.74% WGA on the minority group while Evidential Alignment reaches 84.58%, with majority-group accuracy unchanged. t-SNE embeddings of the test set further show high-uncertainty samples clustering on the minority groups in feature space, confirming that epistemic uncertainty localizes the data the biased model gets wrong.

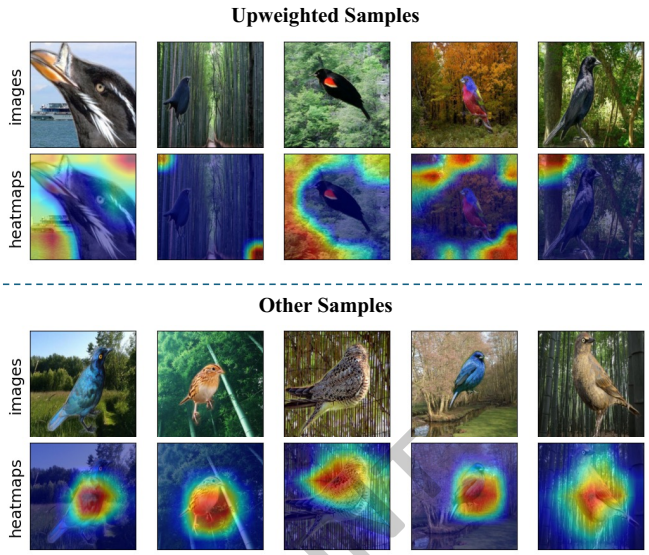


Figure 2: GradCAM visualizations on the Waterbirds dataset. Upweighted (high-uncertainty) samples attend to background regions tied to the spurious attribute, whereas low-uncertainty samples attend to the bird itself.

**Ablations and qualitative analysis.** The full paper ablates the annealing step $\eta$ and proximal coefficient $\beta$, showing that the KL evidence regularization and class-balanced calibration sampling are individually beneficial and combine to give the highest WGA. As Figure 2 illustrates, GradCAM on Waterbirds reveals that the samples our method upweights, those with high epistemic uncertainty, concentrate on background regions tied to the spurious attribute, whereas low-uncertainty samples attend to the bird itself. Quantitative analysis across all four benchmarks shows non-trivial correlation between per-sample uncertainty and the latent group identity, supporting the central claim that epistemic uncertainty is a reliable, annotation-free proxy for spurious-bias-induced failure modes.

## 5 Conclusion

We introduced Evidential Alignment, an uncertainty-guided framework that mitigates spurious correlations without requiring group annotations. By transforming ERM into second-order risk minimization with a Dirichlet parameterization and applying an evidential calibration step that reweights samples by their epistemic uncertainty, our method captures and corrects spurious biases that deterministic auxiliaries miss. Theoretical analysis and extensive experiments across vision and language benchmarks confirm that Evidential Alignment delivers state-of-the-art worst-group accuracy with minimal computational overhead, providing a scalable and principled path toward robust deep learning.

## Acknowledgments

This work is supported in part by the US National Science Foundation under grants 2217071, 2213700, 2106913, and 2008208. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

[Barocas *et al.*, 2023] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[Beery *et al.*, 2018] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pages 456–473, 2018.

[Borkan *et al.*, 2019] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hüllermeier and Waegeman, 2021] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[Khani and Liang, 2021] Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 196–205, 2021.

[Kirichenko *et al.*, 2023] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

[Koh *et al.*, 2021] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[LaBonte *et al.*, 2024] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36, 2024.

[Li *et al.*, 2024] Gaotang Li, Jiarui Liu, and Wei Hu. Bias amplification enhances minority group performance. *Transactions on Machine Learning Research*, 2024.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[Liu *et al.*, 2021] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[Malinin and Gales, 2018] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[Nam *et al.*, 2020] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

[Qiu *et al.*, 2023] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pages 28448–28467. PMLR, 2023.

[Sagawa *et al.*, 2020] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[Stock and Cisse, 2018] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

[Ulmer *et al.*, 2021] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021.

[Vapnik, 1999] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 1999.

[Welinder *et al.*, 2010] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[Williams *et al.*, 2017] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[Yager and Liu, 2010] Ronald R. Yager and Liping Liu. Classic works of the dempster-shafer theory of belief functions. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, 2010.

[Ye *et al.*, 2024] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious Correlations in Machine Learning: A Survey, May 2024. arXiv:2402.12715 [cs].

[Ye *et al.*, 2025] Wenqian Ye, Guangtao Zheng, and Aidong Zhang. Improving group robustness on spurious correlation via evidential alignment. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25)*, 2025.

[Zech *et al.*, 2018] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.