Exploiting What Trained Models Learn for Making Them Robust to Spurious Correlations without Group Annotations

Anonymous authors
Paper under double-blind review

Abstract

Classifiers trained with Empirical Risk Minimization (ERM) often rely on spurious correlations, degrading performance on underrepresented groups and challenging out-of-distribution generalization and fairness. While prior methods aim to address this, many require group annotations for training or validation, limiting their applicability when spurious correlations or group labels are unknown. We demonstrate that what has been learned during ERM training can be utilized to *fully* remove group supervision for both training and model selection. To show this, we design Environment-based Validation and Loss-based Sampling (EVaLS), which uses losses from an ERM-trained model to construct datasets with mitigated group imbalance. EVaLS leverages environment inference to create diverse environments with correlation shifts, enabling model selection without group-annotated validation data. By using worst environment accuracy as a tuning surrogate, EVaLS achieves robust performance across groups through simple last-layer retraining. This fast and effective approach eliminates the need for group annotations, achieving competitive worst-group accuracy and improving robustness to known and unknown spurious correlations.

1 Introduction

Training deep learning models with Empirical Risk Minimization (ERM) risks relying on *spurious correlations*—patterns in the training data that correlate with the target without causal relevance. Learning such shortcuts can reduce accuracy on *minority groups* lacking these patterns (Kirichenko et al., 2023; LaBonte et al., 2023), raising fairness concerns (Hashimoto et al., 2018) and impairing performance. This issue is particularly critical when underrepresented minority groups during training become overrepresented at inference due to subpopulation shifts (Yang et al., 2023b). Ensuring robustness to group shifts and improving *worst-group accuracy* (WGA) is therefore essential for both fairness and reliability in deep learning.

Many studies have proposed solutions to address this challenge. A promising line of research focuses on increasing the contribution of minority groups in the model's training (Liu et al., 2021a; Yang et al., 2023a; Sagawa et al., 2019). A strong assumption that is considered by some previous works is having access to group annotations for training or fully/partially fine-tuning a pretrained model (Nam et al., 2021; Sagawa et al., 2019; Kirichenko et al., 2023). The study by Kirichenko et al. (2023) proposes that retraining the last layer of a model on a dataset that is balanced in terms of group annotation can effectively enhance the model's robustness against shifts in spurious correlation. While these works have shown tremendous robustness performance, their assumption for the availability of the group annotation restricts their usage.

In many real-world applications, the process of labeling samples according to their groups can be prohibitively expensive, and sometimes impractical, especially when all minority groups may not be identifiable beforehand. A widely adopted strategy in these situations involves the indirect inference of various groups, followed by the training of models using a loss function that is balanced across groups (Liu et al., 2021a; Qiu et al., 2023; Nam et al., 2020; Yang et al., 2023b). The loss value of the model, or its alternatives, are popular signals for recognizing minority groups (Liu et al., 2021a; Qiu et al., 2023; Nam et al., 2020; Noohdani et al., 2024). While most of these techniques necessitate full training of a model, Qiu et al. (2023) attempts to adapt the DFR method (Kirichenko et al., 2023) to preserve computational efficiency while simultaneously improving robustness to the group shift. However, this method still requires group annotations of the validation set for the model selection and hyperparameter tuning. Consequently, this constitutes a restrictive assumption when adequate annotations for certain groups are not supplied. It also applies to situations where some shortcut attributes are completely unknown.

In this study, we investigate whether a model trained with standard ERM contains all the information needed to make it robust to spurious correlations. We propose a strategy that mitigates reliance on spurious correlations, eliminating the need for group annotations in both training and retraining stages. Notably, we provide empirical evidence that group annotations are unnecessary even for model selection. Instead, assembling diverse environments reflecting group shifts serves as an effective alternative. Our method, Environment-based Validation and Loss-based Sampling (EVaLS), enhances model robustness against spurious correlations without relying on ground-truth group annotations. In the robustification of trained models, EVaLS is pioneering in its ability to eliminate the need for group annotations at *every phase*, including the model selection step. EVaLS posits that in the absence of group annotations, a set of *environments* showcasing group shifts is sufficient for robustness. This indicates that Worst Environment Accuracy (WEA) could then be utilized for model selection. We observe that spurious correlations cause significant group shifts when using environment inference methods. Consequently, the inferred environments offer a practical mechanism to compare different hyperparameter settings. Figure 2 illustrates the main components of EVaLS.

Aligned with AFR (Qiu et al., 2023) and DFR (Kirichenko et al., 2023), EVaLS offers a significant advantage by not requiring any modifications to the standard ERM training procedure or the original training data. This characteristic is particularly beneficial in enhancing the robustness of ERM-pretrained networks against their potential inherent biases. Specifically, it eliminates the need to retrain the entire model, which may be impractical or infeasible when the original training data is unavailable. While AFR and DFR rely on similar post-hoc approaches, EVaLS demonstrates that even without explicit information about spurious correlations or group annotations, pretrained models inherently contain the signals needed to improve their robustness against spurious correlations.

Our empirical observations support prior research suggesting that loss of a trained model could be a signal for distinguishing samples of minority and majority groups. (Liu et al., 2021a; Qiu et al., 2023; Nam et al., 2020). EVaLS evenly selects from both high-loss and low-loss data to form a balanced dataset for last-layer retraining. Both theoretical insights and empirical results support the effectiveness of this strategy. Comprehensive experiments conducted on spurious correlation benchmarks show that EVaLS achieves accuracy comparable to state-of-the-art methods, despite having no supervision regarding spurious correlations sources. Moreover, when group annotations are accessible solely for model selection, our approach, EVaLS-GL, further improves performance against various distribution shifts, including attribute imbalance and class imbalance. Finally, we introduce two new datasets designed to model real-world scenarios where there are multiple independent shortcuts, but group annotations are not available for some attributes. In this setting, we show that EVaLS, without utilizing group annotations for either attribute, can paradoxically be more effective in preserving the performance of the most underrepresented group.

The main contributions of this paper are as follows:

- We present EVaLS, a simple yet effective post-hoc approach that enhances the robustness of ERM-pretrained models against both known and unknown spurious correlations, without relying on group annotations.
- We provide both theoretical insights and empirical results on how balanced sampling from high-loss and low-loss samples offers a dataset in which the group imbalance is notably mitigated. We further show that worst environment accuracy could serve as a reliable indicator for model selection.
- EVaLS achieves competitive performance in spurious correlation benchmarks without requiring group annotations and delivers state-of-the-art performance when group annotations are available for model selection.
- By introducing two new datasets with multiple spurious attributes and one being unannotated, we show that EVaLS, counterintuitively, improves the robustness of underrepresented groups better than methods relying on group annotation.

2 Preliminaries

2.1 Problem Setting

We consider a supervised learning problem with training set $\mathcal{D}^{\text{Tr}}$, validation set $\mathcal{D}^{\text{Val}}$, and test set $\mathcal{D}^{\text{Te}}$. Each dataset consists of paired samples (x, y), representing the input data and their corresponding label. While conventional settings assume these datasets are sampled from the same distribution, real-world scenarios often involve distribution shifts. Specifically, we address the subpopulation shift problem (Yang et al., 2023b), where data samples belong to groups $\mathcal{G}_i$, each characterized by a shared property. The overall data distribution is given by $p(x, y) = \sum_i \alpha_i p_i(x, y)$, where α_i represents the proportion of group i such that $\sum_i \alpha_i = 1$. In this work, we assume that $\mathcal{D}^{\text{Tr}}$, $\mathcal{D}^{\text{Val}}$, and $\mathcal{D}^{\text{Te}}$ contain the same groups but differ in their mixing coefficients $\{\alpha_i\}$.

Several types of subpopulation shifts have been identified in the literature, including class imbalance, attribute imbalance, and spurious correlation (Yang et al., 2023b) (see Appendix A). In the case of spurious correlation, when a spurious attribute is strongly associated with a label, deep models may rely on it as a shortcut instead of learning the core features. Consequently, model performance degrades on groups lacking this attribute. Our objective is to enhance classifier robustness to spurious attributes by improving performance across all groups.

2.2 Partially Annotated Multiple Spurious Attributes

As previously discussed by Li et al. (2023), when data contains multiple spurious attributes and annotations are only available for some of them, methods that depend on group annotations for training or model selection would make the model robust only to the known spurious attributes. To explore such complex scenarios, we introduce the *Dominoes Colored-MNIST-FashionMNIST (Dominoes-CMF)* (Figure 1(a)) and *CelebA Straight Hair - Smiling - Gender (CelebA-SHSG)* datasets. Drawing inspiration from Pagliardini et al. (2022a) and Arjovsky et al. (2020), Dominoes-CMF merges an image from CIFAR10 (Krizhevsky & Hinton, 2009) at the top with a colored (red or green) MNIST (Deng, 2012) or FashionMNIST (Xiao et al., 2017) image at the bottom. The primary label is derived from the CIFAR10 image, while the bottom part introduces two independent spurious attributes: color (red or green) and style (MNIST or FashionMNIST). Although annotations for shape are provided for training and model selection, color remains an unknown variable until testing. For details on CelebA-SHSG and more details on Dominoes-CMF refer to Appendix F.

In Section 4.1 we show that our approach, which does not rely on the group annotations of the identified group, achieves enhanced robustness to both spurious correlations, outperforming strategies that depend on the known group's information.

3 Environment-based Validation and Loss-based Sampling

In line with the DFR (Kirichenko et al., 2023) approach, we utilize a classifier defined as $f = h_\phi \circ g_\theta$, where g_θ represents a deep neural network serving as a feature extractor, and h_ϕ denotes a linear classifier. The classifier is initially trained with the ERM objective on the training dataset $\mathcal{D}^{\text{Tr}}$. Subsequently, we freeze the feature extractor g_θ and focus solely on retraining the last linear layer h_ϕ using the validation dataset $\mathcal{D}^{\text{Val}}$ as a held-out dataset. This scheme helps us make our method available in settings where $\mathcal{D}^{\text{Tr}}$ is not available, or where repeating the training process is infeasible. We randomly divide the validation set $\mathcal{D}^{\text{Val}}$ into two subsets, $\mathcal{D}^{\text{LL}}$ and $\mathcal{D}^{\text{MS}}$, used for last layer training and model selection, respectively.

In Section 3.1 we explain how to sample a subset of $\mathcal{D}^{\text{LL}}$ that statistically handles the group shifts inherent in the dataset. Building on this, We discuss the theoretical justification for this approach. Here, we leave out the detailed analysis in Appendix E. Finally, in Section 3.2 we describe how $\mathcal{D}^{\text{MS}}$ is divided into different environments that are later used for model selection. The optimal number of selected samples from $\mathcal{D}^{\text{LL}}$ and other hyperparameters is determined based on the worst environment accuracies among environments that are obtained from $\mathcal{D}^{\text{MS}}$. Figure 2 (Appendix B) illustrates the comprehensive workflow of the EVaLS.

3.1 Loss-Based Instance Sampling

Following previous works (Liu et al., 2021a; Nam et al., 2020; Qiu et al., 2023), we use the loss value as an indicator for identifying minority groups. We first evaluate classifier f on samples within $\mathcal{D}^{\mathrm{LL}}$ and choose k samples with the highest and lowest loss values in each class. By combining these $2k$ samples from each class, we construct a balanced set $\mathcal{D}^{\mathrm{Bal}}$, consisting of high-loss and low-loss samples (see Figure 2(c)). $\mathcal{D}^{\mathrm{Bal}}$ is then used for the training of the last layer of the model. We observe that as we choose smaller number of samples with the highest loss, the proportion of minority samples among these samples increases. This suggests that high and low-loss samples could serve as effective representatives of minority and majority groups, respectively (see Figure 3 in Appendix D). However, our approach brings up the question of whether loss-based sampling can successfully construct a *balanced* dataset without introducing spurious correlations. We provide theoretical insights into why this approach may result in group-balanced data in Appendix E.

3.2 Partitioning Validation Set into Environments

Contrary to common assumptions and practices in the field, precise group labels for the validation set are not essential for training models robust to spurious correlations. Our empirical findings, detailed in Section 4, reveal partitioning the validation set into environments that exhibit significant subpopulation shifts can be used for model selection. Under these conditions, the worst environment accuracy (WEA) emerges as a viable metric for selecting the most effective model and hyperparameters.

The *environment* concept in invariant learning refers to data partitions with different distributions. A model with high worst environment accuracy (WEA) is likely to generalize well across test groups. Several methods infer environments with distribution shifts (Creager et al., 2021; Liu et al., 2021b). EIIL (Creager et al., 2021) uses predictions from a trained ERM model to split data into two environments that deviate from the invariant learning principle of Arjovsky et al. (2020), introducing distribution shifts. Initially, an environment inference method (*e.g.* EIIL) is employed to split $\mathcal{D}^{\mathrm{MS}}$ into two environments. Subsequently, each environment is further divided based on sample labels, resulting in $2 \times |\mathcal{Y}|$ environments. To measure the difference between the distribution of environments, we define *group shift* of a class as the absolute difference in the proportion of a minority group between two environments of that class. A higher group shift suggests a more distinct separation between environments. As detailed in the Appendix, environments inferred by EIIL demonstrate an average group shift of 28.7% over datasets with spurious correlation. Further details on environment inference methods and the extent of group shifts in each dataset are provided in the Appendix.

We demonstrate that even more straightforward techniques can be effective to an extent in several cases (See Appendix H.3). These observations underscore that the feature space of a trained model is a valuable resource of information for identifying groups affected by spurious correlations. This supports the logic of previous research that employs clustering (Sohoni et al., 2020) or contrastive methods (Zhang et al., 2021) in this space to differentiate between groups.

4 Experiments

Datasets Our approach, along with other baselines, is evaluated on Waterbirds (Sagawa et al., 2019), CelebA (Liu et al., 2014), UrbanCars (Li et al., 2023), CivilComments (Borkan et al., 2019), and MultiNLI (Williams et al., 2017). As per the study by Yang et al. (2023b), Waterbirds, CelebA, and UrbanCars exhibit spurious correlation. CivilComments has class and attribute imbalance, whereas MultiNLI exhibits attribute imbalance. For additional details, please refer to Appendix G.1.

Baselines We compare EVaLS with six baselines in addition to standard ERM: GroupDRO, DFR (Kirichenko et al., 2023), GroupDRO + EIIL (Creager et al., 2021), JTT (Liu et al., 2021a), ES Disagreement SELF (LaBonte et al., 2023), and AFR (Qiu et al., 2023). GroupDRO + EIIL, JTT, ES Disagreement SELF, and AFR do not rely on group annotations for (re)training but need group labels for model selection, unlike EVaLS. JTT, GroupDRO, and GroupDRO + EIIL require full model training. ES Disagreement and SELF need early-stopped versions during ERM training. DFR, AFR, and EVaLS, however, work entirely post-training without any information from ERM training, making them useful when training data or checkpoints are unavailable or repeating training is costly. More details on the baselines can be found in Appendix G.2

Table 1: Comparison of worst-group accuracy across various methods, including ours, on five datasets. The Group Info column indicates if each method utilizes group labels of the training/validation data, with ✓✓ denoting that group information is employed during both stages. Bold numbers are the highest results overall, while underlined ones are the best among methods that may require group annotation only for model selection. CivilComments is class imbalanced, MultiNLI has imbalanced attributes, and the other three datasets have spurious correlations. The × sign indicates that the dataset is out of the scope of the method. Methods that do not rely on ERM training information are identified with $\star$. Mean and standard deviation are calculated over three runs.

Method	Group Info	Datasets				
	Train/Val	Waterbirds	CelebA	UrbanCars	CivilComments	MultiNLI
GDRO (Sagawa et al., 2019)	✓/✓	91.4	**88.9**	73.1	69.9	**77.7**
DFR* (Kirichenko et al., 2023)	✗/✓✓	$\mathbf{92.9_{\pm 0.2}}$	$88.3_{\pm 1.1}$	$79.6_{\pm 2.2}$	$\mathbf{70.1_{\pm 0.8}}$	$74.7_{\pm 0.7}$
GDRO + EIIL (Creager et al., 2021)	✗/✓	$77.2_{\pm 1}$	$81.7_{\pm 0.8}$	$76.5_{\pm 2.6}$	$67.0_{\pm 2.4}$	$61.2_{\pm 0.5}$
JTT (Liu et al., 2021a)	✗/✓	86.7	81.1	79.5	$\underline{69.3}$	72.6
SELF (LaBonte et al., 2023)	✗/✓	$\underline{91.6_{\pm 1.4}}$	$83.9_{\pm 0.9}$	$83.2_{\pm 0.8}$	$66.0_{\pm 1.7}$	$70.7_{\pm 2.5}$
AFR* (Qiu et al., 2023)	✗/✓	$90.4_{\pm 1.1}$	$82.0_{\pm 0.5}$	$80.2_{\pm 2.0}$	$68.7_{\pm 0.6}$	$73.4_{\pm 0.6}$
EVaLS-GL* (Ours)	✗/✓	$89.4_{\pm 0.3}$	$\underline{84.6_{\pm 1.6}}$	$\underline{\mathbf{83.5_{\pm 1.7}}}$	$68.0_{\pm 0.5}$	$\underline{75.1_{\pm 1.2}}$
EVaLS* (Ours)	✗/✗	$88.4_{\pm 3.1}$	$\underline{85.3_{\pm 0.4}}$	$82.1_{\pm 0.9}$	×	×
ERM	✗/✗	$66.4_{\pm 2.3}$	$47.4_{\pm 2.3}$	$18.67_{\pm 2.0}$	$56.3_{\pm 4.8}$	$64.8_{\pm 1.9}$

Setup Model selection and hyper-parameter fine-tuning are done according to the worst environment (or group if annotations are assumed to be available) accuracy on the validation set. For each dataset, we assess the performance of our model in two cases: fine-tuning the ERM classifier or retraining it. We report results in two settings: (i) EVaLS, which incorporates loss-based instance sampling for training the last layer, and environment inference for model selection. (ii) EVaLS-GL, similar to EVaLS except in using ground-truth group labels for model selection. For more details on the setup, ERM training and last layer training hyperparameters refer to the Appendix.

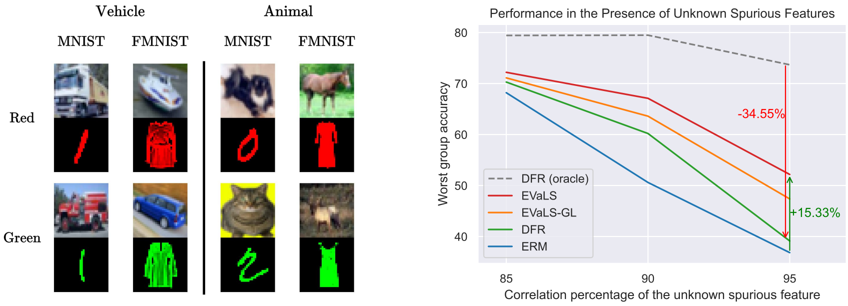
Results The results are shown in Table 1. Overall, our approaches outperform methods that do not require group annotations for (re)training in 2 out of 3 datasets with spurious correlations. Moreover, EVaLS-GL surpasses other methods with a similar level of group supervision on MultiNLI and achieves state-of-the-art performance among all methods on UrbanCars . Furthermore, EVaLS and EVaLS-GL, similar to DFR (Kirichenko et al., 2023) and AFR (Qiu et al., 2023), can be applied to ERM-trained models without needing further information about their training.

The comparison between EVaLS and GroupDRO + EIIL indicates that when environments are available instead of groups, our method, which uses environments solely for model selection and utilizes loss-based sampling, is more effective than GroupDRO, a potent invariant learning method.

Our evaluation of EVaLS is based on the spurious correlation benchmarks. This is because, in other instances of subpopulation shift, the attributes that differ across groups are not predictive of the label, thereby reducing the visibility of these attributes' effects in the model's final layers (Lee et al., 2023). Consequently, EIIL, which depends on output logits for prediction, might not effectively separate the groups. This observation is further supported by our findings related to the degree of group shift between the environments inferred by EIIL for each class in the CivilComments and MultiNLI datasets. The average group shift (defined in the Section 3.2) in the environments of the minority class of CivilComments is only $0.8_{\pm 0.0}\%$. Also, environments associated with Classes 1 and 2 in MultiNLI show only $1.1_{\pm 0.3}\%$ and $1.9_{\pm 1.0}\%$ group shift respectively. More results and ablation studies can be found in the Appendix.

4.1 Mitigating Unknown Spurious Correlations

To assess the performance of our method in scenarios with unknown spurious correlations, we evaluate DFR (Kirichenko et al., 2023), AFR (Qiu et al., 2023), AFR + EIIL (Creager et al., 2021), EVaLS-GL, and EVaLS on the Dominoes-CMF and CelebA-SHSG datasets (Section 2.2). It is worth mentioning that in these two datasets, unlike UrbanCars Li et al. (2023), the spurious correlation in the training set is also present in the validation data (which is used for retraining phase and hyperparameter tuning). All methods are applied on the last layer of ResNet-18 models trained with ERM. We set the spurious correlation of the known attribute to 75% in Dominoes-CMF and 80%



(a) (b)

Figure 1: (a) The Dominoes-CMF dataset, which contains two spurious attributes. (b) Performance on Dominoes-CMF is measured by worst-group accuracy across varying levels of correlation between the target label and the unknown spurious attribute (color). The performance gap between EVaLS and EVaLS-GL with lower group supervision compared to DFR (Kirichenko et al., 2023) increases with higher correlations.

in CelebA-SHSG, and conduct experiments for various amounts of unknown spurious correlation. During model selection, we calculate the worst-group accuracy on the validation set considering only the label of the known shortcut, *i.e.*, the lowest accuracy among the four groups based on the combination of the target label and the single known shortcut label. However, the final results on the test data are based on the worst-group accuracies, with groups defined by the labels of both spurious attributes. The results on Dominoes-CMF are shown in Figure 1(b). Note that EVaLS utilizes information about neither known nor unknown spurious attributes.

Our results reveal that methods using group labels mitigate reliance on the known shortcut but not necessarily on the unknown one. DFR shows a significant drop in performance when it relies on a single known spurious attribute for grouping, compared to the oracle that uses both grouping attributes. EVaLS-GL mitigates this issue using loss-based sampling, but EVaLS even outperforms EVaLS-GL. Combining loss-based sampling for last-layer training and environment-based model selection results in a completely group-annotation-free method in a setting with unknown spurious correlations, and successfully re-weights features to perform well with respect to multiple spurious attributes. The effectiveness of removing the supervision on validation group labels in reducing the worst-group accuracy is also observable in the comparison between AFR and AFR+EIIL. It is also evident that increasing unknown spurious correlations results in a larger gap between the performance of EVaLS and DFR. The complete results on Dominoes-CMF and Celeba-SHSG are in Appendix F.

5 Discussion

This study presents EVaLS, a novel approach to improve robustness to spurious correlations with zero group annotation. EVaLS uses loss-based sampling to create a balanced training dataset and employs environment inference to guide model selection. Unlike existing methods that rely on explicit group supervision, EVaLS operates fully annotation-free, making it applicable to real-world scenarios where spurious correlations are unknown or difficult to annotate. We also explore situations with multiple spurious correlations, some of which are unknown. In this context, we introduce Dominoes-CMF and CelebA-SHSG datasets, where two factors are spuriously correlated with the label, but only one is identified. EVaLS—despite operating without group supervision—outperforms methods that depend on partial group supervision in this setting. Additionally, EVaLS-GL, which uses group labels only for model selection, surpasses state-of-the-art approaches that require group labels for training or evaluation.

EVaLS enhances model fairness and can be applied to ERM-pretrained models without prior bias knowledge. However, its effectiveness may decline on small datasets, and it is limited to cases where spurious correlations exist. Future research could strengthen its theoretical foundations and extend environment inference techniques to address other subpopulation shifts, such as attribute and class imbalance.

REFERENCES

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=b9PoimzZFJ`.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

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*, pp. 491–500, 2019.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2189–2200. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/creager21a.html`.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. *ArXiv*, abs/2306.11957, 2023. URL `https://api.semanticscholar.org/CorpusID:259211935`.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Zb6c8A-Fghk`.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/koh21a.html`.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/krueger21a.html`.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=kshC3NOP6h`.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=APuPRxjHvZ`.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20071–20082, June 2023.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021a. URL `https://proceedings.mlr.press/v139/liu21f.html`.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6804–6814. PMLR, 18–24 Jul 2021b. URL `https://proceedings.mlr.press/v139/liu21h.html`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2014. URL `https://api.semanticscholar.org/CorpusID:459456`.

Yuzhen Mao, Zhun Deng, Huaxiu Yao, Ting Ye, Kenji Kawaguchi, and James Zou. Last-layer fairness fine-tuning is simple and effective for neural networks. *arXiv preprint arXiv:2304.03935*, 2023.

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.

Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2021.

Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, HamidReza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. *CoRR*, abs/2402.18919, 2024. doi: 10.48550/ARXIV.2402.18919. URL `https://doi.org/10.48550/arXiv.2402.18919`.

Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. Agree to disagree: Diversity through disagreement for better transferability. In *The Eleventh International Conference on Learning Representations*, 2022a.

Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint*, arXiv:2202.04414, 2022b.

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*, pp. 28448–28467. PMLR, 2023.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2180–2188, 2022.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *ArXiv*, abs/2108.13624, 2021. URL `https://api.semanticscholar.org/CorpusID:237364121`.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. Group robust classification without any group information. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 56553–56575. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/b0d9ceb3d11d013e55da201d2a2c07b2-Paper-Conference.pdf`.

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL `http://arxiv.org/abs/1708.07747`. cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. *ArXiv*, abs/2305.18761, 2023a. URL `https://api.semanticscholar.org/CorpusID:258967752`.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, 2023b.

Michael Zhang, Nimit Sharad Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. URL `https://openreview.net/forum?id=Q41kl_DwS3Y`.

A Related Work

Several kinds of subpopulation shifts are defined in the literature, including class imbalance, attribute imbalance, and spurious correlation (Yang et al., 2023b). Class imbalance occurs when there is a difference between the proportion of samples from each class, while attribute imbalance arises when instances with a certain attribute are underrepresented in the training data, even though this attribute may not necessarily be a reliable predictor of the label. On the other hand, spurious correlation occurs when various groups are differentiated by spurious attributes that are partially predictive and correlated with class labels but are causally irrelevant. More precisely, we can consider a set of spurious attributes $\mathcal{S}$ that partition the data into $|\mathcal{S}| \times |\mathcal{Y}|$ groups. When a spurious attribute is strongly correlated with a label, deep models may use it as a shortcut instead of core features. This is followed by a decrease in the model's performance on groups that do not have this attribute.

Given a class, the group containing samples with correlated spurious attributes is referred to as *majority* group of that class, while the other groups are called the *minority* groups. As an example, in the Waterbirds dataset (Sagawa et al., 2019), for which the task is to classify images of birds into landbird and waterbird, there are spurious attributes $\{$*water background*, *land background*$\}$. Each background is spuriously correlated with its associated label, decompose the data into two majority groups *waterbird on water background*, and *landbird on land background*, and two minority groups *waterbird on land background* and *landbird on water background*.

Robustness to spurious correlation is a critical concern across various machine learning subfields. It is a form of out-of-distribution generalization (Shen et al., 2021) where the distribution shift arises from the disproportionate representation of minority groups—those instances that are devoid of the correlated spurious patterns associated with their labels (Yang et al., 2023b). The issue of spurious correlation also intersects with the discourse on fairness in machine learning (Seo et al., 2022; Mao et al., 2023).

Past studies have proposed a range of strategies to mitigate the models' reliance on spurious correlation. Broadly speaking, these methods can be categorized according to the degree of supervision they require regarding group labels.

Invariant learning (IL) methods (Arjovsky et al., 2020; Krueger et al., 2021; Rame et al., 2022) operate under the assumption of having access to a collection of environments that comprise group shift. By imposing invariant conditions on these environments, IL methods strive to create classifiers robust against group-sensitive features. IRM (Arjovsky et al., 2020) is designed to learn a feature extractor, which, when utilized, guarantees the existence of a classifier that would be optimal in all training environments. VREx (Krueger et al., 2021) aims to decrease the risk variance among different training environments. PGI (Ahmed et al., 2021) works by minimizing the distance between the expected softmax distribution of labels, conditioned on inputs across both majority and minority environments. Lastly, Fishr (Rame et al., 2022) focuses on bringing the variance of risk gradients closer together across different training environments. For scenarios that the environments are not available, environment inference methods (Creager et al., 2021; Liu et al., 2021b) are used to obtain a set of environments. Creager et al. (2021) introduce environment inference for invariant learning (EIIL), which tries to partition samples into two groups such that the objective of IRM (Arjovsky et al., 2020) is maximized. HRM (Liu et al., 2021b) aims to optimize both an environment inference module and an invariant prediction module jointly, with the goal of achieving an invariant predictor.

When group annotations are accessible, various methods leverage this information to equalize the impact of different groups on the model's loss. The Group Distributionally Robust Optimization (GDRO) approach (Sagawa et al., 2019), for instance, focuses on optimizing the loss for the worst-performing group during training. Kirichenko et al. (2023) has shown that models can still learn and extract core data features even in the presence high spurious correlation. Consequently, They suggest that retraining just the last layer of a model initially trained with Empirical Risk Minimization (ERM) can effectively reduce reliance on spurious correlation for predicting class labels. This method, termed Deep Feature Re-weighting (DFR), has been validated as not only highly effective but also significantly more efficient than earlier techniques that necessitated retraining the full model (Nam et al., 2021; Sagawa et al., 2019). However, availability of group annotations is considered a serious restrictive assumption.

Several recent studies have endeavored to enhance model robustness against spurious correlation, even in the absence of group annotations (Liu et al., 2021a; Zhang et al., 2021; Qiu et al., 2023; LaBonte et al., 2023; Yang et al., 2023a; Tsirigotis et al., 2023). Liu et al. (2021a) introduce a two-stage method that involves training a model using ERM for a number of epochs before retraining it to give more weight to misclassified samples. The study by Zhang et al. (2021) employs the same two-stage training process, but with a twist for the second stage: they utilize contrastive methods. The goal is to bring samples from the same class but with divergent predictions closer in the feature space, while simultaneously increasing the separation between samples from different classes that have similar predictions. Another method, known as automatic feature reweighting (AFR) (Qiu et al., 2023), reweights the last layer of an ERM-pretrained model to favor samples that the original model was less accurate on. LaBonte et al. (2023) refine the last layer of an ERM-trained model through class-balanced finetuning, identifying challenging data points by comparing the classifier's predictions with those of an early-stopped version. While these methods have significantly reduced the reliance on group annotations, they still required for validation and model selection. This remains a constraint, particularly when the spurious correlation is completely unknown.

To make a trained model robust to subpopulation shifts with zero group annotations, LaBonte et al. (2023) have recently demonstrated that class-balanced retraining of a model pretrained with ERM can effectively improve the worst-group accuracy (WGA) for certain datasets. While this method effectively reduces the impact of class imbalance, it fails in datasets with spurious correlations.

B Method Overview

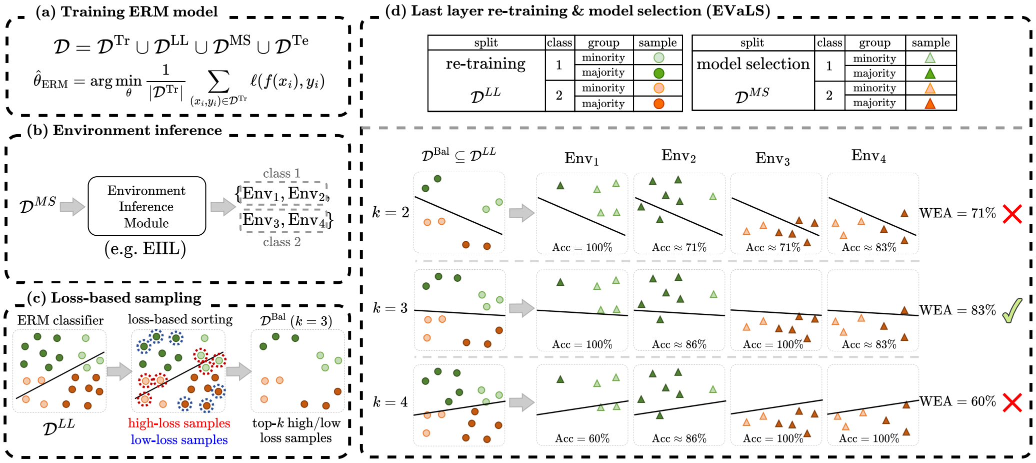

(a) Training ERM model
$\mathcal{D} = \mathcal{D}^{\mathrm{Tr}} \cup \mathcal{D}^{\mathrm{LL}} \cup \mathcal{D}^{\mathrm{MS}} \cup \mathcal{D}^{\mathrm{Te}}$
$\hat{\theta}_{\mathrm{ERM}} = \arg\min_{\theta} \frac{1}{|\mathcal{D}^{\mathrm{Tr}}|} \sum_{(x_i,y_i)\in\mathcal{D}^{\mathrm{Tr}}} \ell(f(x_i), y_i)$
(b) Environment inference
$\mathcal{D}^{MS}$
Environment Inference Module
(e.g. EIIL)
class 1
$\{\mathrm{Env}_1, \mathrm{Env}_2,$
$\mathrm{Env}_3, \mathrm{Env}_4\}$
class 2
(c) Loss-based sampling
ERM classifier
loss-based sorting
$\mathcal{D}^{\mathrm{Bal}}$ $(k = 3)$
$\mathcal{D}^{LL}$
high-loss samples
low-loss samples
top-k high/low loss samples
(d) Last layer re-training & model selection (EVaLS)
split | class | group | sample
re-training $\mathcal{D}^{LL}$ | 1 | minority | majority | 2 | minority | majority
model selection $\mathcal{D}^{MS}$ | 1 | minority | majority | 2 | minority | majority
$\mathcal{D}^{\mathrm{Bal}} \subseteq \mathcal{D}^{LL}$
Env_1
Env_2
Env_3
Env_4
$k = 2$
Acc = 100%
Acc ≈ 71%
Acc ≈ 71%
Acc ≈ 83%
WEA = 71%
$k = 3$
Acc = 100%
Acc ≈ 86%
Acc = 100%
Acc ≈ 83%
WEA = 83%
$k = 4$
Acc = 60%
Acc ≈ 86%
Acc = 100%
Acc = 100%
WEA = 60%

Figure 2: Overview of the proposed approach. (a) We randomly split the dataset $\mathcal{D}$ into $\mathcal{D}^{\mathrm{Tr}}$, $\mathcal{D}^{\mathrm{MS}}$, $\mathcal{D}^{\mathrm{LL}}$ and $\mathcal{D}^{\mathrm{Te}}$. We train the initial classifier on $\mathcal{D}^{\mathrm{Tr}}$ with empirical risk minimization (ERM). (b) An environment inference method is utilized to infer diverse environments for each class of $\mathcal{D}^{\mathrm{MS}}$. (c) We evaluate $\mathcal{D}^{\mathrm{LL}}$ samples on the initial ERM classifier and sort high-loss and low-loss samples of each class for loss-based sampling. (d) Finally, we perform last-layer retraining on the loss-based selected samples $\mathcal{D}^{\mathrm{Bal}}$. Each retraining setting (*e.g.* different k for loss-based sampling) is validated based on the worst accuracy of the inferred environments. Note that majority and minority groups are shown with dark and light colors for better visualization, but are not known in our setting.

Algorithm 1 EVaLS

```
Input: Held-out dataset $\mathcal{D}^{\mathrm{Val}}$, ERM-trained model $f_{\mathrm{ERM}}$, maximum $k$ value $k_{\max}$
Output: Optimal number of samples $k^*$, best model $f^*$, best performance wea$^*$
$(\mathcal{D}^{\mathrm{LL}}, \mathcal{D}^{\mathrm{MS}}) \leftarrow$ splitDataset$(\mathcal{D}^{\mathrm{Val}})$ {Split the held-out dataset}
Envs$[y] \leftarrow$ inferEnvs$(\mathcal{D}^{\mathrm{MS}})[y]$   $\forall y \in \mathcal{Y}$
sortedSamples$[y] \leftarrow$ sortByLoss$(f_{\mathrm{ERM}}, \mathcal{D}^{\mathrm{LL}}[y])$ $\forall y \in \mathcal{Y}$
Initialize wea$^* \leftarrow 0$, $k^* \leftarrow 0$, $f^* \leftarrow$ None
for $k = 1$ to $k_{\max}$ do
    highLossSamples$[y] \leftarrow$ sortedSamples$[y][:k]$   $\forall y \in \mathcal{Y}$ {Select top-$k$ high-loss samples}
    lowLossSamples$[y] \leftarrow$ sortedSamples$[y][-k:]$   $\forall y \in \mathcal{Y}$ {Select top-$k$ low-loss samples}
    $\mathcal{D}^{\mathrm{Bal}} \leftarrow$ {highLossSamples, lowLossSamples} {Combine samples}
    $f \leftarrow$ retrainLastLayer$(\mathcal{D}^{\mathrm{Bal}})$ {Retrain the last layer with combined samples}
    wea $\leftarrow$ evaluateWEA$(f, \mathrm{Envs})$ {Evaluate the retrained model}
    if wea > wea$^*$ then
        wea$^* \leftarrow$ wea, $f^* \leftarrow f$, $k^* \leftarrow k$ {Record the best configuration}
    end if
end for
Return: $k^*$, wea$^*$, $f^*$
```

C Environment Inference for Invariant Learning

Consider the training dataset $\mathcal{D}^{\mathrm{Tr}} = \{(x^{(i)}, y^{(i)}) | x^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}\}$, where $\mathcal{X}$ and $\mathcal{Y}$ represent the input and output spaces, respectively. This dataset can be partitioned into different environments $\mathcal{E}^{tr} = \{e_1, ..., e_n\}$, such that for any $i \neq j$, the data distribution in e_i and e_j differs. The objective of invariant learning is to train a predictor that performs consistently across all environments in $\mathcal{E}^{tr}$. Under certain conditions, this predictor is also expected to perform well on e^{tst}, a test environment with a distribution distinct from the training data. Invariant Risk Minimization (IRM) (Arjovsky et al.,

Table 2: The average and variation percentage (%)(across 3 seeds) of group shift between the inferred environments using EIIL (Creager et al., 2021) for each class, which is the absolute difference between the proportion of a minority group in the two environments of a class. Higher group shift indicates better separation of environments. In most cases, a significant group shift is observed between the inferred environments.

Class No.	Dataset		
	Waterbirds	CelebA	UrbanCars
0	$16.6_{\pm 0.7}$	$3.6_{\pm 0.2}$	$17.7_{\pm 1.2}, 23.5_{\pm 0.1}, 62.1_{\pm 1.9}$
1	$50.5_{\pm 0.3}$	$14.1_{\pm 0.9}$	$40.7_{\pm 7.9}, 13.8_{\pm 0.1}, 19.2_{\pm 3.9}$

2020) approaches this problem by learning a feature extractor $\Phi(.)$ such that a classifier $\omega(.)$ exists, where $\omega \circ \Phi(.)$ performs consistently across all training environments. The practical implementation of the IRM objective is to minimize

$$\sum_{e \in \mathcal{E}^{tr}} R^e(\Phi) + \lambda ||\nabla_{\bar{\omega}} R^e(\bar{\omega} \circ \Phi)||^2, \tag{1}$$

where $\bar{\omega}$ is a constant scalar with a value of 1.0, λ is a hyperparameter, and $R^e(f) = \mathbb{E}_{(x,y)\sim p_e}[l(f(x), y)]$ is referred to as the risk on environment e.

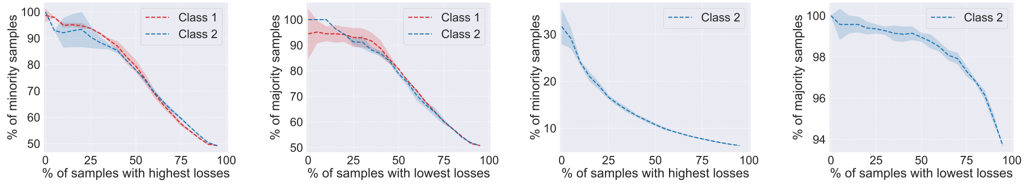
In real-world scenarios, training environments might not always be available. To address this, Environment Inference for Invariant Learning (EIIL) (Creager et al., 2021) partitions samples into two environments in a way that maximizes the objective in Eq 1.

During the training phase, the EIIL algorithm replaces the hard assignment of environments to samples with a soft assignment $\mathbf{q}_i(e) = p(e|(x^{(i)}, y^{(i)}))$, where $\mathbf{q}_i$ is learnable. Consequently, the relaxed version of the risk function is defined as $\tilde{R}^e(\Phi) = \frac{1}{N}\sum_i^N \mathbf{q}_i(e)[l(\Phi(x^{(i)}), y^{(i)})]$. Given a model Φ that has been trained with ERM on the dataset, EIIL optimizes

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} ||\nabla_{\bar{\omega}} \tilde{R}^e(\bar{\omega} \circ \Phi)||. \tag{2}$$

As discussed in Creager et al. (2021), using a biased base model Φ could lead to environments exhibiting varying degrees of spurious correlation. During the inference phase, the soft assignment is converted to a hard assignment. The average group shift between the inferred environments using EIIL is illustrated in Table 2.

D Groups Distribution Across Loss Thresholds



(a) Waterbirds

(b) CelebA

Figure 3: The percentage of samples with the highest (lowest) losses across various thresholds that belong to the minority (majority) group within different classes in $\mathcal{D}^{\text{LL}}$ for (a) the Waterbirds and (b) CelebA datasets. Minority group samples are more prevalent among high-loss samples, while majority group samples dominate the low-loss areas. Note that in the CelebA dataset, only the "blond hair" class includes a minority group. The error bars are calculated across three ERM models.

E Theoretical Analysis on Loss-Based Sampling

In this section, we establish a formal description of loss-based sampling for balanced dataset creation and then prove it. We thoroughly analyze the close relationship between the availability of the balanced dataset and the gap between spurious features of minority and majority groups.

E.1 Feasibility Of Loss-based Group Balancing

Consider a binary classification problem with a cross-entropy loss function. Let logits be denoted as L. Because loss is a monotonic function of logits, the tails of the distribution of loss across samples are equivalent to that of the logits in each class. We assume that in feature space (output of g_θ) samples from the minority and majority of a class are derived from Gaussian distributions $\mathcal{N}(h_{\text{min}}, \Sigma_{\text{min}})$ and $\mathcal{N}(h_{\text{maj}}, \Sigma_{\text{maj}})$, respectively. Before diving into the group balance problem we initially show that the distribution of minority and majority samples in the logit space (output of h_ϕ) are Gaussian too.

Lemma E.1. *[Gaussain Distribution of Logits] Considering a Gaussian distribution $Z \sim \mathcal{N}(h, \Sigma)$ in feature space and $W \in \mathbb{R}^d$, then the distribution of logits is as follows: $L = \langle W, Z\rangle \sim \mathcal{N}\big(Wh,\, \|W\|^2_\Sigma\big)$.*

Proof. Let $Z \sim \mathcal{N}(h, \Sigma)$.

Consider $L = \langle W, Z\rangle = W^T Z$, where $W \in \mathbb{R}^d$. L is a linear combination of jointly gaussian random variables which makes it an univariate gaussian random variable.

To find the distribution of L, we need to determine its mean and variance.

1. **Mean of** L

$$\begin{aligned}\mathbb{E}[L] = \mathbb{E}[\langle W, Z\rangle] = \mathbb{E}[W^T Z] = \\ W^T \mathbb{E}[Z] = W^T h = \langle W, h\rangle.\end{aligned}$$

Therefore, the mean of L is Wh.

2. **Variance of** L:

The variance of L can be computed using the properties of covariance. Recall that if $Z \sim \mathcal{N}(h, \Sigma)$, then the covariance matrix of Z is Σ.

The variance of the linear combination $L = W^T Z$ is given by:

$$\mathrm{Var}(L) = \mathrm{Var}(W^T Z) = W^T \Sigma W = \|W\|_\Sigma^2 ,$$

where $\|W\|_\Sigma$ denotes the Mahalanobis norm of W.

Thus, we have proved that if $Z \sim \mathcal{N}(h, \Sigma)$, then the logits $L = \langle W, Z \rangle$ follow the distribution $\mathcal{N}(Wh, \|W\|_\Sigma^2)$. □

From now on, we consider $\mathcal{N}(\mu_{\text{min}}, \sigma_{\text{min}}^2)$ and $\mathcal{N}(\mu_{\text{maj}}, \sigma_{\text{maj}}^2)$ as the distribution of minority and majority samples in logits space.

Next, we prove the more formal version of the main proposition **??**, which describes the existence of a balanced dataset, only after we define a key concept, *proportional density difference* (illustrated in figure 4) to outline our proof.

Definition E.1 (Proportional Density Difference)**.** *For any interval $I = (a, b]$ and a mixture distribution $\varepsilon P_1(x) + (1 - \varepsilon)P_2(x)$, proportional density difference is defined by the difference of accumulation of two component distributions in the interval I and is denoted by $\Delta_\varepsilon P_{mixture}(I)$.*

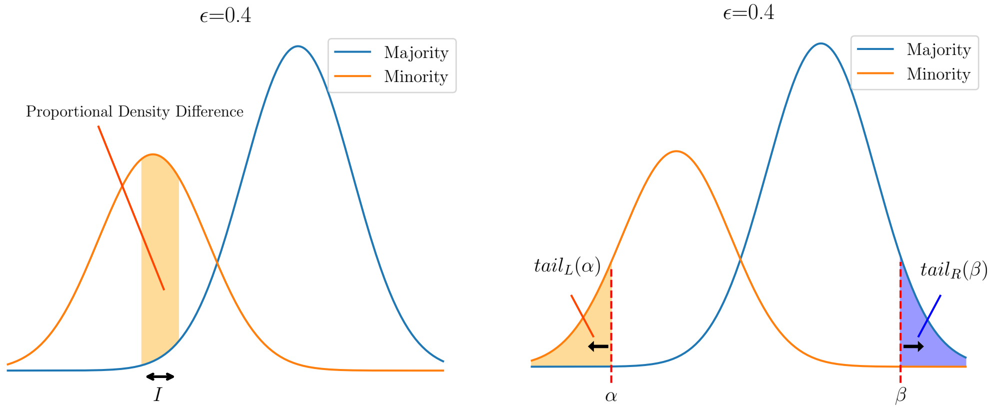
$$\Delta_\varepsilon P_{mixture}(I) \triangleq \varepsilon P_1\big(x \in I\big) - (1 - \varepsilon)P_2\big(x \in I\big)$$

Definition E.2 (Tail Proportional Density Difference)**.** *For a mixture distribution $\varepsilon P_1(x) + (1 - \varepsilon)P_2(x)$, we define $tail_L(\alpha)$ as $\Delta_\varepsilon P_{mixture}\Big((-\infty, \alpha]\Big)$ and $tail_R(\beta)$ as $-\Delta_\varepsilon P_{mixture}\Big((\beta, +\infty)\Big)$.*

Corollary E.1.

$$tail_L(\alpha) = \varepsilon F^1(\alpha) - (1 - \varepsilon)F^2(\alpha)$$
$$tail_R(\beta) = (1 - \varepsilon)\big[1 - F^2(\beta)\big] - \varepsilon\big[1 - F^1(\beta)\big]$$

where F^1 and F^2 are CDF of two component distributions.



(a)

(b)

Figure 4: (a) Illustration of proportion density difference E.1, (b) equation of $tail_L(\alpha) = tail_R(\beta)$ at E.2.

Proposition E.1. *[Feasiblity Of Loss-based Group Balancing] Suppose that L is derived from the mixture of two distributions $\mathcal{N}(\mu_{min}, \sigma^2_{min})$ and $\mathcal{N}(\mu_{maj}, \sigma^2_{maj})$ with proportion of ε and $1-\varepsilon$, respectively, where $\varepsilon \leq \frac{1}{2}$. There exists α and β such that restricting L to the α-left and β-right tails of its distribution results in a group-balanced distribution if and only if (i)*

$$\sigma_{min} \geq \sigma_{maj}, \tag{3}$$

or (ii)

$$tail_L(\frac{-B+\sqrt{\Delta}}{2A}) > 0 \tag{4}$$

and

$$\epsilon \geq sigmoid\Bigg(- \frac{(\mu_{maj} - \mu_{min})^2}{2(\sigma^2_{maj} - \sigma^2_{min})} - \log\Big(\frac{\sigma_{maj}}{\sigma_{min}}\Big)\Bigg). \tag{5}$$

where $A = \Big(\frac{1}{2\sigma^2_{maj}} - \frac{1}{2\sigma^2_{min}}\Big)$, $B = \Big(\frac{\mu_{min}}{\sigma^2_{min}} - \frac{\mu_{maj}}{\sigma^2_{maj}}\Big)$ *and* $\Delta = \frac{(\mu_{min}-\mu_{maj})^2}{\sigma^2_{min}\sigma^2_{maj}} - 4\Big[\log\Big(\frac{\sigma_{maj}}{\sigma_{min}}\Big) + \log\Big(\frac{\epsilon}{1-\epsilon}\Big)\Big]\Big[\frac{1}{2\sigma^2_{maj}} - \frac{1}{2\sigma^2_{min}}\Big]$.

A quick analysis of conditions in our method's setting. If $\sigma_{\text{max}} > \sigma_{\text{min}}$, condition 5 suggests that for a given degree of spurious correlation ϵ and variations $\sigma_{\text{maj}}, \sigma_{\text{min}}$, an essential prerequisite for the efficacy of loss-based sampling is a sufficiently large disparity between the mean distributions of minority and majority samples, denoted by $\|\mu_{\text{maj}} - \mu_{\text{min}}\|^2$. This indicates that the groups should be distinctly separable in the logits space. In other case, $\sigma_{\text{min}} > \sigma_{\text{maj}}$, group-balance is always feasible.

Proof outline

Our proof proceeds with three steps. First, we reformulate the theorem as an equality of left- and right-tail proportional distribution differences. In other words, we show that the more mass the minority distribution has on one tail, the more mass the majority distribution must have on the other tail. Afterward, supposing $\mu_{\text{min}} < \mu_{\text{maj}}$ WLOG , we propose a proper range for β values on the right tail. We show that when $\sigma_{\text{maj}} \leq \sigma_{\text{min}}$, values for α trivially exist that can overcome the imbalance between the two distributions. In the last step, for the case in which the variance of the majority is higher than the minority, we discuss a necessary and sufficient condition for the existence of α and β based on the left-tail proportional density difference using the properties of its derivative with respect to α.

Step 1 *Reformulating the problem based on proportional distribution difference.*

We introduce a utility random variable *Logit Value Tier* as T, which is defined as a function of a random variable L.

$$T_{\alpha,\beta} = \begin{cases} High & \text{if } L \geq \beta \\ Mid & \text{if } \alpha < L < \beta \\ Low & \text{if } L \leq \alpha \end{cases} \tag{6}$$

We can rewrite the problem in formal form as finding an α and β which satisfies the following equation:

$$P\Big(g = \text{min}\Big|T_{\alpha,\beta} \neq\ Mid\Big) = P\Big(g = \text{maj}\Big|T_{\alpha,\beta} \neq Mid\Big) \tag{7}$$

Equation 5 now can be rewritten to a more suitable form:

$$P\Big(g = \text{min}\Big|T_{\alpha,\beta} \neq Mid\Big) = P\Big(g = \text{maj}\Big|T_{\alpha,\beta} \neq Mid\Big) \tag{8}$$

$$\begin{aligned}
&\iff \frac{P\Big(T_{\alpha,\beta} \neq Mid\Big|g = \text{min}\Big)P(g = \text{min})}{P\Big(T_{\alpha,\beta} \neq Mid\Big)} = \\
&\frac{P\Big(T_{\alpha,\beta} \neq Mid|g = \text{maj}\Big)P(g = \text{maj})}{P\Big(T_{\alpha,\beta} \neq Mid\Big)}
\end{aligned} \tag{9}$$

$$\begin{aligned}
&\iff P\Big(T_{\alpha,\beta} \neq Mid\Big|g = \text{min}\Big)P(g = \text{min}) = \\
&P\Big(T_{\alpha,\beta} \neq Mid\Big|g = \text{maj}\Big)P(g = \text{maj})
\end{aligned} \tag{10}$$

$$\begin{aligned}
&\iff \varepsilon P\Big(T_{\alpha,\beta} \neq Mid\Big|g = \text{min}\Big) = \\
&(1-\varepsilon)P\Big(T_{\alpha,\beta} \neq Mid\Big|g = \text{maj}\Big)
\end{aligned} \tag{11}$$

$$\begin{aligned}
&\iff \varepsilon\Big[P\Big(T_{\alpha,\beta} = Low\Big|g = \text{min}\Big)+ \\
&P\Big(T_{\alpha,\beta} = High\Big|g = \text{min}\Big)\Big] = \\
&(1-\varepsilon)\Big[P\Big(T_{\alpha,\beta} = Low\Big|g = \text{maj}\Big)+ \\
&P\Big(T_{\alpha,\beta} = High\Big|g = \text{maj}\Big)\Big]
\end{aligned} \tag{12}$$

$$\begin{aligned}
&\iff \varepsilon\Big[P\Big(L \leq \alpha\Big|g = \text{min}\Big)+ \\
&P\Big(L \geq \beta\Big|g = \text{min}\Big)\Big] = \\
&(1-\varepsilon)\Big[P\Big(L \leq \alpha\Big|g = \text{maj}\Big)+ \\
&P\Big(L \geq \beta\Big|g = \text{maj}\Big)\Big]
\end{aligned} \tag{13}$$

$$\begin{aligned}
&\iff \varepsilon\Big[F^{\text{min}}(\alpha) + \Big(1 - F^{\text{min}}(\beta)\Big)\Big] = \\
&(1-\varepsilon)\Big[F^{\text{maj}}(\alpha) + \Big(1 - F^{\text{maj}}(\beta)\Big)\Big]
\end{aligned} \tag{14}$$

$$\begin{aligned}
&\iff \varepsilon F^{\text{min}}(\alpha) - (1-\varepsilon)F^{\text{maj}}(\alpha) = \\
&(1-\varepsilon)\Big[1 - F^{\text{maj}}(\beta)\Big] - \varepsilon\Big[1 - F^{\text{min}}(\beta)\Big]
\end{aligned} \tag{15}$$

We can see the left side of equation 15 is just a function of $alpha$. The same goes for the right side of the equation which is a function of β.

Rewriting the left side of the equation as $tail_L(\alpha)$ and right side as $tail_R(\beta)$, the problem is now reduced to finding an α and β that satisfies

$$tail_L(\alpha) = tail_R(\beta) \tag{16}$$

which is shown in figure 4.

Before reaching out to step two we discuss the properties of $tail_L$ and $tail_R$ in Lemma E.2.

Lemma E.2. *$tail_L(\alpha)$ and $tail_R(\beta)$ are continuous functions and $\lim_{\alpha\to-\infty} tail_L(\alpha) = 0$, $\lim_{\alpha\to+\infty} tail_L(\alpha) = 2\varepsilon - 1 < 0$, $\lim_{\beta\to+\infty} tail_R(\beta) = 0$ and $\lim_{\beta\to-\infty} tail_R(\beta) = 1 - 2\varepsilon > 0$.*

Proof. Simply proved by the definition of *tail* functions and properties of CDF. □

Step 2 *Solving the equation 16 for simple cases.*

Lemma E.3. *$tail_R(\mu_{maj}) > \frac{1}{2} - \varepsilon \geq 0$*

Proof.

$$
\begin{aligned}
& tail_R(\mu_{\text{maj}}) \\
&= (1-\varepsilon)\Big[1 - F^{\text{maj}}(\mu_{\text{maj}})\Big] - \varepsilon\Big[1 - F^{\text{min}}(\mu_{\text{maj}})\Big] && (17) \\
&= (1-\varepsilon)\Big[1 - \phi(0)\Big] - \varepsilon\Big[1 - \phi(\frac{\mu_{\text{maj}} - \mu_{\text{min}}}{\sigma_{\text{min}}})\Big] && (18) \\
&> \frac{(1-\varepsilon)}{2} - \varepsilon\big(1 - \frac{1}{2}\big) = \frac{1-2\varepsilon}{2} = \frac{1}{2} - \varepsilon && (19)
\end{aligned}
$$

□

Corollary E.2. *Because $tail_R$ is continuous and $\lim_{\beta\to+\infty} tail_R(\beta) = 0$, based on the intermediate value theorem, any value between zero and $\frac{(1-2\varepsilon)}{2}$ is obtainable by selecting a β in $[\mu_2, +\infty)$.*

According to the previous corollary E.2 finding a positive $tail_L(\alpha)$ will satisfy our need. to find a suitable point, we employ derivatives and properties of relative PDFs to maximize $tail_L(\alpha)$ and find a positive value.

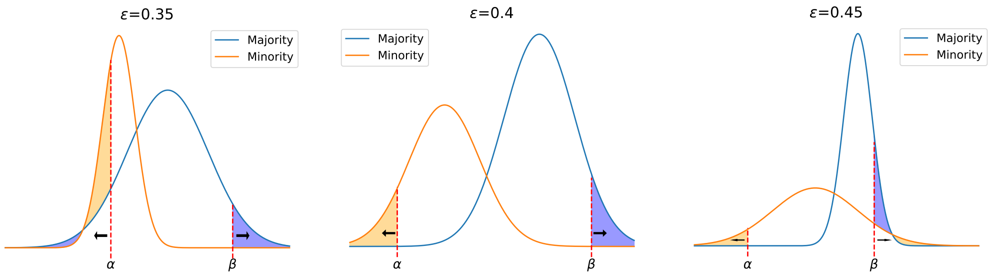
$$
\begin{aligned}
\frac{\mathrm{d}tail_L(\alpha)}{\mathrm{d}\alpha} &= \varepsilon f^{\text{min}}(\alpha) - (1-\varepsilon) f^{\text{maj}}(\alpha) && (20) \\
&= \varepsilon f^{\text{maj}}(\alpha)\Big[\frac{f^{\text{min}}(\alpha)}{f^{\text{maj}}(\alpha)} - \frac{1-\varepsilon}{\varepsilon}\Big] && (21)
\end{aligned}
$$

The term $[\frac{f^{\text{min}}(\alpha)}{f^{\text{maj}}(\alpha)} - \frac{1-\varepsilon}{\varepsilon}]$ has the same sign with derivative of $tail_L(\alpha)$, also it's roots are critical points of $tail_L$, analyzing characteristics of $\log \frac{f^{\text{min}}(\alpha)}{f^{\text{maj}}(\alpha)}$ is the key insight to find a proper α value.

$$
\log f^{\text{min}}(\alpha) - \log f^{\text{maj}}(\alpha) = \log\left(\frac{1-\epsilon}{\epsilon}\right)
$$

$$
\begin{aligned}
\Rightarrow \log\left(\frac{\sigma_{\text{maj}}}{\sigma_{\text{min}}}\right) - \log\left(\frac{1-\epsilon}{\epsilon}\right) - \frac{(\alpha - \mu_{\text{min}})^2}{2\sigma_{\text{min}}^2} \\
+ \frac{(\alpha - \mu_{\text{maj}})^2}{2\sigma_{\text{maj}}^2} = 0
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow \left(\frac{1}{2\sigma_{\text{maj}}^2} - \frac{1}{2\sigma_{\text{min}}^2}\right)\alpha^2 + \\
&\left(\frac{\mu_{\text{min}}}{\sigma_{\text{min}}^2} - \frac{\mu_{\text{maj}}}{\sigma_{\text{maj}}^2}\right)\alpha + \\
&\Bigg[\frac{\mu_{\text{maj}}^2}{2\sigma_{\text{maj}}^2} - \frac{\mu_{\text{min}}^2}{2\sigma_{\text{min}}^2} + \\
&\quad \log\left(\frac{\sigma_{\text{maj}}}{\sigma_{\text{min}}}\right) + \log\left(\frac{\epsilon}{1-\epsilon}\right)\Bigg] = 0
\end{aligned}
$$



(a) (b) (c)

Figure 5: Tail thresholds for three cases: (a) minority group variance is less than majority ($\sigma_{\text{min}} < \sigma_{\text{maj}}$), (b) the variance of two groups are equal ($\sigma_{\text{min}} = \sigma_{\text{maj}}$) and (c) the variance of the minority group is more than majority ($\sigma_{\text{min}} > \sigma_{\text{maj}}$).

Because $\lim_{\alpha \to -\infty} tail_L(\alpha) = 0$ and also $\lim_{\beta \to +\infty} tail_R(\beta) < 0$, to have a positive $tail_L(\alpha)$, we need to have an interval which $\frac{\mathrm{d}tail_L(\alpha)}{\mathrm{d}\alpha}$ is positive. For a second degree polynomial like $ax^2 + bx + c$ to have positive value, either $a \geq 0$ or $\Delta > 0$, in our case a is $\left(\frac{1}{\sigma_{\text{maj}}^2} - \frac{1}{\sigma_{\text{min}}^2}\right)$. if $\sigma_{\text{min}} \geq \sigma_{\text{maj}}$ then $a \geq 0$ and the minority CDF function will dominate the majority CDF function in the left-side tail and by choosing a negative number with big enough absolute value for alpha and $tail_L(\alpha)$ will be positive.

Step 3 *Solving equation 16 for special case $\sigma_{min} < \sigma_{maj}$* In case of $\sigma_{\text{min}} \leq \sigma_{\text{maj}}$, having $\Delta > 0$ is a necessary condition, also derivative of $tail_L(\alpha)$ is only positive in $(\frac{-b-\sqrt{\Delta}}{2a}, \frac{-b+\sqrt{\Delta}}{2a})$ so the maximum of $tail_L$ is either in $-\infty$ or in $\frac{-b+\sqrt{\Delta}}{2a}$. Having $tail_L(\frac{-b+\sqrt{\Delta}}{2a}) > 0$ next to $\Delta > 0$ condition, would be the necessary and also sufficient in this case.

$$B^2 = \frac{\mu_{\text{min}}^2}{\sigma_{\text{min}}^4} + \frac{\mu_{\text{maj}}^2}{\sigma_{\text{maj}}^4} - 2\frac{\mu_{\text{maj}}\mu_{\text{min}}}{\sigma_{\text{maj}}^2\sigma_{\text{min}}^2}$$

$$\begin{aligned} 4AC =& \frac{\mu_{\text{min}}^2}{\sigma_{\text{min}}^4} - \frac{\mu_{\text{min}}^2}{\sigma_{\text{maj}}^2\sigma_{\text{min}}^2} - \\ & \frac{\mu_{\text{maj}}^2}{\sigma_{\text{maj}}^2\sigma_{\text{min}}^2} + \frac{\mu_{\text{maj}}^2}{\sigma_{\text{maj}}^4} + \\ & 4\left[\log\left(\frac{\sigma_{\text{maj}}}{\sigma_{\text{min}}}\right) + \log\left(\frac{\epsilon}{1-\epsilon}\right)\right] \times \\ & \left[\frac{1}{2\sigma_{\text{maj}}^2} - \frac{1}{2\sigma_{\text{min}}^2}\right] \end{aligned}$$

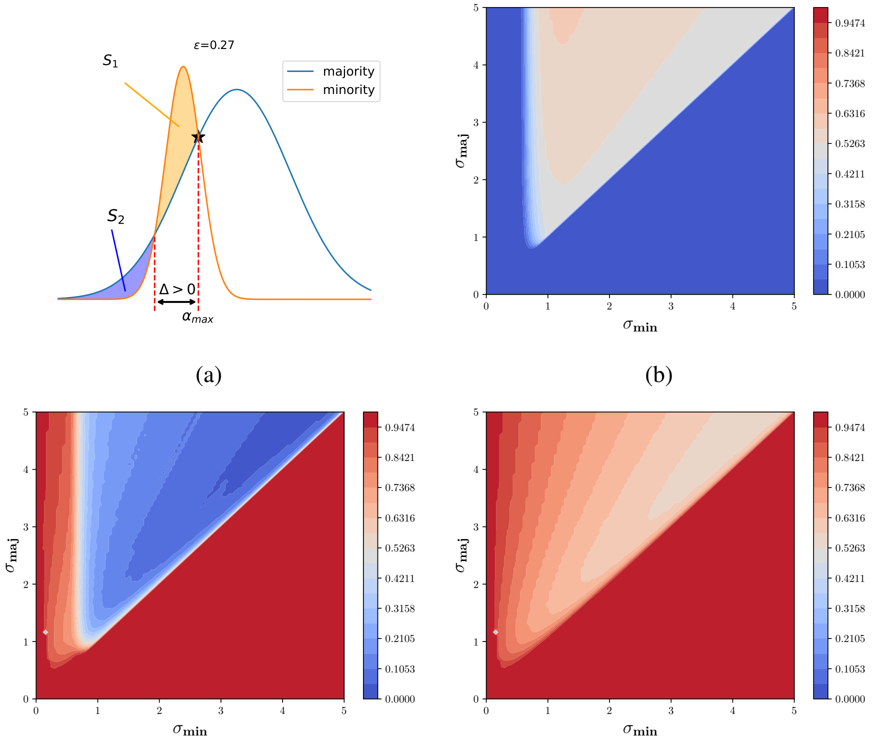
$$\begin{aligned} \Delta =& \frac{(\mu_{\text{min}} - \mu_{\text{maj}})^2}{\sigma_{\text{min}}^2\sigma_{\text{maj}}^2} - \\ & 4\left[\log\left(\frac{\sigma_{\text{maj}}}{\sigma_{\text{min}}}\right) + \log\left(\frac{\epsilon}{1-\epsilon}\right)\right] \times \\ & \left[\frac{1}{2\sigma_{\text{maj}}^2} - \frac{1}{2\sigma_{\text{min}}^2}\right] \geq 0 \end{aligned}$$

$$\begin{aligned} \iff & (\mu_{\text{min}} - \mu_{\text{maj}})^2 \\ \geq 2\left[\log\left(\frac{1-\epsilon}{\epsilon}\right) - \log\left(\frac{\sigma_{\text{maj}}}{\sigma_{\text{min}}}\right)\right] & \left[\sigma_{\text{maj}}^2 - \sigma_{\text{min}}^2\right] \end{aligned}$$

$$\iff \epsilon \geq \text{sigmoid}\left(-\frac{(\mu_{\text{maj}} - \mu_{\text{min}})^2}{2(\sigma_{\text{maj}}^2 - \sigma_{\text{min}}^2)} - \log\left(\frac{\sigma_{\text{maj}}}{\sigma_{\text{min}}}\right)\right)$$

Next, we investigate properties of the conditions of the proposition E.1 in case of $\sigma_{\text{maj}} < \sigma_{\text{min}}$. Schematic interpretation of these conditions is presented in figure 6.

- As equation 5 indicates, the minority group is not allowed to be too underrepresented. This especially has a direct relation with the difference of means. The more mean values of groups are different, the more imbalance can be mitigated through loss-based sampling. Mean value difference is especially affected by the spurious correlation, it escalates as the model relies on spurious correlation and also when the spurious features between groups are too different.
- On the other hand condition 4 is more complex and doesn't have a simple closed form, we analytically describe its behaviors by fixating the means and calculating the valid values for ε. As the results show in figure 6, most of ε are feasible in for $\sigma_{\text{min}} < \Delta\mu$ as we can see the possible region declines with an increase of σ_{min} and valid ε values cease to exist.



(a)

(b)

(d)

(c)

Figure 6: (a) Conditions if $\sigma_{\text{min}} > \sigma_{\text{maj}}$, (b), (c), (d) minimum, maximum and interval length of feasible ε values across $(\sigma_{\text{min}}, \sigma_{\text{maj}})$ field for $\mu_{\text{min}} = 0$, $\mu_{\text{maj}} = 1$.

E.2 Practical Justification

As shown in Table 3, the standard deviation (σ) of the minority group is consistently greater than that of the majority group across all analyzed datasets. Consequently, condition (i) (Eq. 3) of Proposition E.1 is satisfied. Therefore, we theoretically expect the existence of properly balanced left and right tails.

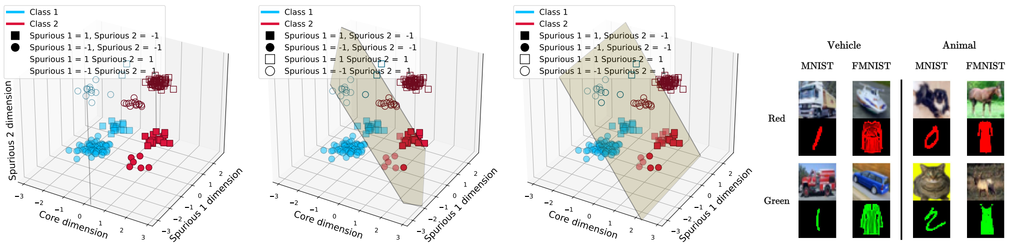
E.3 Hyperparameter Tuning and Practical Considerations

Although the parameters α and β are theoretically established under certain conditions, their actual values remain undetermined. Therefore, validation data is essential to identify the appropriate tails. For practicality and simplicity, we assume an equal number k of samples for both tails and explore this count (high- and low-loss samples) from a predefined set of values. By leveraging the worst environment accuracy on validation data after last-layer retraining, as detailed in Section 3.2, we identify the potential candidate that ensures uniform accuracy across all environments.

Table 3: Means, standard deviations (STD), and Earth Mover's Distance across WaterBirds and CelebA datasets.

	Waterbirds				**CelebA**	
	Class 1		**Class 2**		**Class 2**	
	Min	**Maj**	**Min**	**Maj**	**Min**	**Maj**
Mean (μ)	-6.77	-19.17	2.55	11.39	-1.02	6.42
STD (σ)	6.31	6.23	6.97	4.75	7.64	6.48
Earth Mover's Distance	12.40		8.84		7.43	

F Partially Annotated Multiple Spurious Correlations



(a) (b) (c) (d)

Figure 7: (a) If all spurious attributes in a dataset are known, they can be utilized to fit a classifier that captures the essential attributes. (b) In the absence of knowledge about all spurious attributes, the model would rely on them for classification, leading to incorrect classification of minority samples. (c) If some spurious attribute is unknown (Spurious 2), the model becomes robust only to the known spurious correlations (Spurious 1), but still underperforms on minority samples. (d) The Dominoes-CMF dataset, which contains two spurious attributes.

As mentioned in Section 2.2, when data contains multiple spurious attributes and annotations are only available for one of them, models that depend on group annotation for training or model selection would make the model robust only to the known spurious attributes. The illustrations in Figure 7(a-c) depict the outlined scenario. A classifier trained using ERM is dependent on both spurious features (Figure 7(b)). Yet, achieving robustness against one spurious correlation (Figure 7(c)), does not ensure robustness against both (Figure 7(a)).

F.1 Dominoes-Colored-MNIST-FashionMNIST

Dominoes-Colored-MNIST-FashionMNIST (Dominoes-CMF) is a synthetic dataset. We adopt a similar approach to previous works (Pagliardini et al., 2022b; Shah et al., 2020; Kirichenko et al., 2023) using a modified version of the *Dominoes* binary classification dataset. This dataset consists of images with the top half showing CIFAR-10 images (Krizhevsky & Hinton, 2009), divided into two meaningful classes: vehicles (airplane, car, ship, truck) and animals (cat, dog, horse, deer). The bottom half displays either MNIST (Deng, 2012) images from classes $\{0-3\}$ or Fashion-MNIST (Xiao et al., 2017) images from classes $\{\text{T-shirt}, \text{Dress}, \text{Coat}, \text{Shirt}\}$. The complex feature (top half) serves as the core feature and the simple feature (bottom half) is linearly separable and correlated with the class label at 75%. Furthermore, inspired by the approaches in Zhang et al. (2021); Arjovsky et al. (2020), we intentionally introduce an additional spurious attribute by artificially coloring a subset of images as follows: for three different datasets, 85%, 90%, and 95% of the images in the bottom half of class c_1 are randomly assigned a red color in each respective dataset, while 15%, 10%, and 5% of the images are assigned a green color, respectively. The same procedure is applied inversely for class c_2.

See Tables 4, 5, and 6 for more details about the dataset statistics and a comparison of the performance of different methods on the dataset.

F.2 CELEBA-SHSG

CelebA-Straight Hair- Smiling- Gender (CelebA-SHSG) is a subset of the original CelebA (Liu et al., 2014), where the label *"Straight Hair"* is correlated with the attributes of smiling and being female. The *"Straight Hair"* attribute is considered as the label, *"Smiling"* as the known spurious attribute, and *"gender"* as the unknown spurious attribute. Average accuracies and Worst-group accuracies (WGA) are reported in Table 9 among 8 groups (all binary combinations of the label and spurious attributes). We set the spurious correlation of the *"Smiling"* attribute to 80% and design three datasets with 85%, 90% and 95% of correlation between the *"gender"* attribute and the label (similar to the Dominoes-CMF experiments). Spurious correlations are imposed by subsampling from the original CelebA dataset. See Tables 7, 8, and 9 for more details about the dataset statistics and a comparison of the performance of different methods on the dataset.

Table 4: *Dominoes-CMF* Dataset Statistics for 85%, 90%, and 95% Correlation

Top part		**Bottom Part** (85% **Corr.**)		**Bottom Part** (90% **Corr.**)		**Bottom Part** (95% **Corr.**)	
CIFAR-10 Class	**Color**	**MNIST**	**Fashion-MNIST**	**MNIST**	**Fashion-MNIST**	**MNIST**	**Fashion-MNIST**
c_1 (Vehicle)	Red	12,750	4,250	13,500	4,500	14,250	4,750
	Green	2,250	750	1,500	500	750	250
c_2 (Animal)	Red	750	2,250	500	1,500	250	750
	Green	4,250	12,750	4,500	13,500	4,750	14,250
Total		40,000		40,000		40,000	

Table 5: ERM Accuracies on *Dominoes-CMF* Dataset. The mean and standard deviation are reported based on three runs with different seeds.

Top part		**Bottom Part** (85% **Corr.**)		**Bottom Part** (90% **Corr.**)		**Bottom Part** (95% **Corr.**)	
CIFAR-10 Class	**Color**	**MNIST**	**Fashion-MNIST**	**MNIST**	**Fashion-MNIST**	**MNIST**	**Fashion-MNIST**
c_1 (Vehicle)	Red	$98.53_{\pm 0.01}\%$	$95.61_{\pm 1.1}\%$	$99.2_{\pm 0.01}\%$	$95.2_{\pm 1.1}\%$	$99.63_{\pm 0.01}\%$	$98.11_{\pm 1.1}\%$
	Green	$89.33_{\pm 2.4}\%$	$68.57_{\pm 0.5}\%$	$84.5_{\pm 2.4}\%$	$54.7_{\pm 0.5}\%$	$63.1_{\pm 1.4}\%$	$36.84_{\pm 0.5}\%$
c_2 (Animal)	Red	$68.28_{\pm 2.6}\%$	$86.18_{\pm 2.4}\%$	$56.8_{\pm 5.6}\%$	$86.7_{\pm 2.4}\%$	$39.13_{\pm 1.6}\%$	$68.53_{\pm 2.4}\%$
	Green	$93.97_{\pm 0.5}\%$	$98.36_{\pm 0.2}\%$	$96.2_{\pm 0.5}\%$	$99.3_{\pm 0.2}\%$	$97.92_{\pm 0.5}\%$	$99.25_{\pm 0.2}\%$

Table 6: A Comparison of ERM, DFR, EVaLS, and EVaLS-GL on the Dominoes-CMF with different spurious correlations for the unknown feature. Both the worst and average of test group accuracies are presented. The mean and standard deviation are calculated based on runs with three distinct seeds.

	85% Corr.		90% Corr.		95% Corr.	
Method	Worst	Average	Worst	Average	Worst	Average
ERM	$68.3_{\pm 1.5}$	$97.1_{\pm 0.5}$	$50.6_{\pm 1.0}$	$96.1_{\pm 0.0}$	$36.8_{\pm 2.0}$	$95.4_{\pm 1.0}$
DFR (Oracle)	$79.4_{\pm 0.8}$	$93.4_{\pm 1.1}$	$78.5_{\pm 1.2}$	$92.1_{\pm 0.6}$	$73.7_{\pm 1.5}$	$90.3_{\pm 0.7}$
DFR	$70.7_{\pm 0.5}$	$86.2_{\pm 0.6}$	$60.2_{\pm 1.2}$	$84.6_{\pm 0.4}$	$42.7_{\pm 2.7}$	$81.5_{\pm 1.2}$
AFR	$65.7_{\pm 0.2}$	$94.2_{\pm 0.8}$	$54.2_{\pm 0.2}$	$94.9\pm_{2.1}$	$40.3_{\pm 0.5}$	$95.9_{\pm 1.2}$
AFR + EIIL	$69.1_{\pm 0.1}$	$92_{\pm 1.3}$	$61.5_{\pm 0.2}$	$92.1_{\pm 1.9}$	$40.4_{\pm 0.1}$	$92.9_{\pm 1.5}$
EVaLS-GL	$70.1_{\pm 2.9}$	$82.5_{\pm 1.8}$	$63.6_{\pm 1.3}$	$78.7_{\pm 1.5}$	$48.5_{\pm 0.8}$	$77.0_{\pm 2.0}$
EVaLS	$\mathbf{73.0_{\pm 4.8}}$	$81.5_{\pm 1.8}$	$\mathbf{67.1_{\pm 4.2}}$	$78.6_{\pm 2.0}$	$\mathbf{51.2_{\pm 1.4}}$	$77.5_{\pm 2.5}$

Table 7: *CelebA-SHSG* Dataset Statistics for 85%, 90%, and 95% Correlation of the Known Spurious Feature.

Label		**85% Corr.**		**90% Corr.**		**95% Corr.**	
	Gender	**Smiling = 0**	**Smiling = 1**	**Smiling = 0**	**Smiling = 1**	**Smiling = 0**	**Smiling = 1**
"Straight Hair"= 0	Female	1676	419	1055	264	500	125
	Male	9500	2375	9500	2375	9500	2375
"Straight Hair"= 1	Female	2375	9500	2375	9500	2375	9500
	Male	419	1676	264	1055	125	500
Total		27,940		26,388		25,000	

Table 8: ERM Accuracies on *CelebA-SHSG* Dataset. The mean and standard deviation are reported based on three runs with different seeds.

label		85% **Corr.**		90% **Corr.**		95% **Corr.**	
	Gender	**Smiling = 0**	**Smiling = 1**	**Smiling = 0**	**Smiling = 1**	**Smiling = 0**	**Smiling = 1**
"Straight Hair"= 0	Female	$71.2_{\pm 0.6}$	$47.4_{\pm 1.1}$	$67.1_{\pm 0.7}$	$46.4_{\pm 0.6}$	$58.8_{\pm 1.5}$	$39.4_{\pm 1.9}$
	Male	$94.1_{\pm 0.6}$	$76.8_{\pm 1.0}$	$95.0_{\pm 0.0}$	$80.5_{\pm 0.4}$	$96.5_{\pm 0.7}$	$88.0_{\pm 1.1}$
"Straight Hair"= 1	Female	$84.1_{\pm 0.4}$	$96.0_{\pm 0.4}$	$87.2_{\pm 0.4}$	$95.5_{\pm 0.4}$	$91.0_{\pm 0.5}$	$97.0_{\pm 0.2}$
	Male	$28.3_{\pm 0.6}$	$61.6_{\pm 0.8}$	$23.9_{\pm 1.5}$	$55.8_{\pm 2.2}$	$15.6_{\pm 2.6}$	$37.5_{\pm 3.3}$

G Experimental Details

G.1 Datasets

Waterbirds (Sagawa et al., 2019) The dataset comprises images of diverse bird species, classified into two categories: waterbirds and landbirds. Each image features a bird set against a backdrop of either water or land. Interestingly, the background scene acts as a spurious feature in this classification task. Waterbirds are primarily shown against water backgrounds, and landbirds against land backgrounds. Consequently, waterbirds on water and landbirds on land form the minority groups in the training data. It's important to note that the validation dataset for waterbirds is group-balanced, meaning birds from each class are equally represented against both water and land backgrounds. This dataset is mainly categorized as a spurious correlation dataset.

CelebA (Liu et al., 2014) is a widely used dataset in image classification tasks, featuring annotations for 40 binary facial attributes such as hair color, gender, and age. Hair color classification is particularly prominent in literature focusing on spurious correlation robustness. Notably, gender serves as a spurious attribute within this dataset, where a significant majority 94% of individuals with blond hair are women, while men with blond hair represent a minority group. In addition to spurious correlation in the class of blond hair, this dataset also exhibits class imbalance.

MultiNLI (Williams et al., 2017) dataset involves a text classification task focused on determining the relationship between pairs of sentences: contradiction, entailment, or neutral. Sentences containing negation words such as "no" or "never" are underrepresented in all three classes, inducing attribute imbalance in the dataset. Figure **??** illustrates the distinct behavior of this dataset compared to other datasets that contain spurious attributes.

CivilComments (Borkan et al., 2019) dataset, as part of the WILDS benchmark, involves a text classification task focused on labeling online comments as either "toxic" or "not toxic". Each comment is associated with 8 attributes, including gender (male, female), sexual orientation (LGBTQ), race (black, white), and religion (Christian, Muslim, or other), based on whether these characteristics are mentioned in the comment. While there is a small attribute imbalance in the dataset, it can categorized into datasets with class imbalance. The detailed proportion of each attribute in each class is described in Table 12. In this paper, we use the implementation of the dataset by the `WILDS` package (Koh et al., 2021).

Table 9: A Comparison of ERM, DFR, DFR (Oracle), AFR, AFR+EIIL, EVaLS, and EVaLS-GL on the CelebA-SHSG with different spurious correlations for the unknown feature. Both the worst and average of test group accuracies are presented. The mean and standard deviation are calculated based on runs with three distinct seeds.

Method	85% Corr.		90% Corr.		95% Corr.	
	Worst	Average	Worst	Average	Worst	Average
ERM	$28.3_{\pm 0.6}$	$68.2_{\pm 0.6}$	$23.9_{\pm 1.5}$	$67.3_{\pm 0.3}$	$15.6_{\pm 2.6}$	$63.5_{\pm 0.8}$
DFR (Oracle)	$63.1_{\pm 0.9}$	$71.7_{\pm 1.2}$	$59.2_{\pm 1.9}$	$70.0_{\pm 1.1}$	$58.4_{\pm 5.0}$	$67.7_{\pm 1.5}$
DFR	$27.2_{\pm 2.2}$	$67.7_{\pm 0.2}$	$18.9_{\pm 0.7}$	$64.9_{\pm 0.3}$	$12.3_{\pm 1.6}$	$60.1_{\pm 0.3}$
AFR	$28.1_{\pm 0.4}$	$68.0_{\pm 0.4}$	$24.3_{\pm 2.1}$	$65.7_{\pm 0.0}$	$15.7_{\pm 2.6}$	$63.1_{\pm 0.0}$
AFR + EIIL	$41.3_{\pm 5.7}$	$63.2_{\pm 5.1}$	$36.3_{\pm 4.5}$	$69.8_{\pm 0.0}$	$45.0_{\pm 5.3}$	$63.2_{\pm 0.0}$
EVaLS-GL	$30.5_{\pm 5.2}$	$68.6_{\pm 2.3}$	$26.3_{\pm 6.4}$	$67.4_{\pm 1.0}$	$19.3_{\pm 3.2}$	$61.6_{\pm 3.4}$
EVaLS	$\mathbf{45.2_{\pm 2.9}}$	$59.5_{\pm 2.7}$	$\mathbf{44.9_{\pm 3.1}}$	$62.7_{\pm 1.8}$	$\mathbf{45.7_{\pm 2.2}}$	$64.4_{\pm 1.8}$

Table 10: A comparison of the various methods, ours included, on spurious correlation datasets. The Group Info column indicates if each method utilizes group labels of the training/validation data, with ✓✓ denoting that group information is employed during both the training and validation stages. Both the average test accuracy and worst test group accuracy are reported. The mean and standard deviation are calculated over three runs with different seeds. The numbers in bold represent the highest results among all methods, while the underlined numbers represent the best results among methods that may not require group annotation in the training phase.

Method	Group Info	Waterbirds		CelebA		UrbanCars	
	Train/Val	Worst	Average	Worst	Average	Worst	Average
GDRO (Sagawa et al., 2019)	✓/✓	91.4	93.5	**88.9**	92.9	73.1	$84.2_{\pm 1.3}$
DFR (Kirichenko et al., 2023)	✗/✓✓	$\mathbf{92.9_{\pm 0.2}}$	$94.2_{\pm 0.4}$	$88.3_{\pm 1.1}$	$91.3_{\pm 0.3}$	$79.6_{\pm 2.22}$	$87.5_{\pm 0.6}$
GDRO + EIIL (Creager et al., 2021)	✗/✓	$77.2_{\pm 1}$	$\underline{\mathbf{96.5_{\pm 0.2}}}$	$81.7_{\pm 0.8}$	$85.7_{\pm 0.1}$	$76.5_{\pm 2.6}$	$85.4_{\pm 2.1}$
JTT (Liu et al., 2021a)	✗/✓	86.7	93.3	81.1	88.0	79.5	86.3
SELF (LaBonte et al., 2023)	✗/✓	$\underline{91.6_{\pm 1.4}}$	$93.6_{\pm 1.1}$	$83.9_{\pm 0.9}$	$91.7_{\pm 0.4}$	$83.2_{\pm 0.8}$	$\underline{\mathbf{90.0_{\pm 0.5}}}$
AFR (Qiu et al., 2023)	✗/✓	$90.4_{\pm 1.1}$	$94.2_{1.2}$	$82.0_{\pm 0.5}$	$91.3_{\pm 0.3}$	$80.2_{\pm 2.0}$	$87.1_{\pm 1.2}$
EVaLS-GL (Ours)	✗/✓	$89.4_{\pm 0.3}$	$95.1_{\pm 0.3}$	$\underline{84.6_{\pm 1.6}}$	$91.1_{\pm 0.6}$	$\underline{\mathbf{83.5}_{\pm 1.7}}$	$88.3_{\pm 0.9}$
ERM	✗/✗	$66.4_{\pm 2.3}$	$90.3_{\pm 0.5}$	$47.4_{\pm 2.3}$	$\underline{\mathbf{95.5_{\pm 0.0}}}$	$18.67_{\pm 2.01}$	$76.5_{\pm 4.6}$
EVaLS (Ours)	✗/✗	$88.4_{\pm 3.1}$	$94.1_{\pm 0.1}$	$\underline{85.3_{\pm 0.4}}$	$89.4_{\pm 0.5}$	$82.1_{\pm 0.9}$	$88.1_{\pm 0.9}$

UrbanCars (Li et al., 2023) is an image classification dataset with multiple shortcuts. Each image in the dataset consists of a car in the center of the image on a natural scene background, with another object to the right of the image. Images are labeled *Urban* or *City* according to the type of car present

Table 11: A comparison of the various methods, ours included, on CivilComments and MultiNLI. The Group Info column indicates if each method utilizes group labels of the training/validation data, with ✓✓ denoting that group information is employed during both the training and validation stages. Both the average test accuracy and worst test group accuracy are reported. The mean and standard deviation are calculated over three runs with different seeds. The numbers in bold represent the highest results among all methods, while the underlined numbers represent the best results among methods that may not require group annotation in the training phase.

Method	Group Info	CivilComments		MultiNLI	
	Train/Val	Worst	Average	Worst	Average
GDRO (Sagawa et al., 2019)	✓/✓	**69.9**	88.9	**77.7**	81.4
DFR (Kirichenko et al., 2023)	✗/✓✓	$70.1_{\pm 0.8}$	$87.2_{\pm 0.3}$	$74.7_{\pm 0.7}$	$82.1_{\pm 0.2}$
GDRO + EIIL (Creager et al., 2021)	✗/✓	$67.0_{\pm 2.4}$	$90.5_{\pm 0.2}$	$61.2_{\pm 0.5}$	$79.4_{\pm 0.2}$
JTT (Liu et al., 2021a)	✗/✓	$\underline{69.3}$	91.1	72.6	78.6
SELF (LaBonte et al., 2023)	✗/✓	$65.9_{\pm 1.7}$	$89.7_{\pm 0.6}$	$70.7_{\pm 2.5}$	$81.2_{\pm 0.7}$
AFR (Qiu et al., 2023)	✗/✓	$68.7_{\pm 0.6}$	$89.8_{\pm 0.6}$	$73.4_{\pm 0.6}$	$81.4_{\pm 0.2}$
EVaLS-GL (Ours)	✗/✓	$68.0_{\pm 0.5}$	$89.2_{\pm 0.3}$	$\underline{75.1_{\pm 1.2}}$	$81.6_{\pm 0.2}$
ERM	✗/✗	$56.3_{\pm 4.8}$	$\underline{\mathbf{92.0_{\pm 0.0}}}$	$64.8_{\pm 1.9}$	$\underline{\mathbf{82.6_{\pm 0.0}}}$

Table 12: Proportion of attributes in each class for CivilComments dataset.

Toxicity (Class)	Male	Female	LGBTQ	Christian	Muslim	Other Religions	Black	White
0	0.11	0.12	0.03	0.10	0.05	0.02	0.03	0.05
1	0.14	0.15	0.08	0.08	0.10	0.03	0.1	0.14

in the center. However, each of the backgrounds and the additional objects is highly correlated with the label. While the test set consists of 8 environments based on combinations of the core and two spurious patterns, the training and validation set consist of four groups, based on combinations of the label and only one of the shortcuts.

G.2 Baselines

In this study, we compare EVaLS with six baselines in addition to standard ERM **GroupDRO** (Sagawa et al., 2019) trains a model on the data with the objective of minimizing its average loss on the minority samples, requiring group labels for both training and validation. **DFR** (Kirichenko et al., 2023) first trains a model with ERM, then retrains the last linear classifier on a group-balanced subset of validation or held-out held-out training data. While DFR reduces the number of group-annotated samples, it still requires group labels in the training phase. **GroupDRO + EIIL** (Creager et al., 2021) infers environments of the training set and trains a model with GroupDRO on the inferred environments. **JTT** (Liu et al., 2021a) first trains a model with ERM on the dataset, and then retrains it by upweighting the samples misclassified by the ERM model. **ES Disagreement SELF** (LaBonte et al., 2023) fine-tunes the last layer of the ERM-trained model on samples selected based on output differences between the ERM-trained model and its early-stopped version. **AFR** (Qiu et al., 2023) trains a model with standard ERM, and retrains the classifier on a weighted held-out data. The weights assigned to retraining samples are determined by the probability that the ERM-pretrained model assigns to the ground-truth label, leading to an increased weighting of samples from minority groups.

G.3 Complete Results

The complete results on Waterbirds, CelebA, and UrbanCars, in addition to complete results on CivilComments and MultiNLI are reported in Tables 10 and 11 respectively. The reported results for GroupDRO, DFR, JTT, and AFR except those for the UrbanCars are taken from Qiu et al. (2023). For EIIL+Group DRO, the results for Waterbirds, CelebA, and CivilComments are reported from Zhang et al. (2021). The results of SELF on CelebA and MultiNLI are reported from LaBonte et al. (2023). We report only the worst-group accuracy of methods in Table 1. The average accuracies are documented in the Appendix. The Group Info column shows whether group annotation is required for training or model selection. Methods that do not require information regarding ERM training (such as training data or checkpoints) are identified with a $star$ in the table.

Also, the results of our method and DFR are shown in Table 6.

G.4 Training Details

ERM Similar to all the works mentioned in Section G.2, we use ResNet-50 (He et al., 2016) pretrained on ImageNet (Russakovsky et al., 2015) for image classification tasks. We used random crop and random horizontal flip as data augmentation, similar to Kirichenko et al. (2023). For a fair comparison with the baselines, we did not employ any data augmentation techniques in the process of retraining the last layer of the model. For the CivilComments and MultiNLI, we use pretrained BERT (Devlin et al., 2019) and crop sentences to 220 tokens length. In EvaLS, we use the implementation of EIIL by `spuco` package (Joshi et al., 2023) for environments inference on the model selection set with 20000 steps, SGD optimizer, and learning rate 10^{-2} for all datasets.

For Waterbirds and CelebA, we utilize the ResNet50 checkpoints available in the GitHub repository of Kirichenko et al. (2023) as our base model. We use the ResNet-50 architecture provided by the `torchvision` package. In the case of CivilComments and MultiNLI, we adopt a similar approach to Kirichenko et al. (2023), using `BertForSequenceClassification.from_pretrained('bert-base-uncased', ...)` from the `transformers` package.

The model is trained using the AdamW optimizer with a learning rate of 10^{-5}, weight decay of 10^{-4}, and a batch size of 16 for a total of 5 epochs.

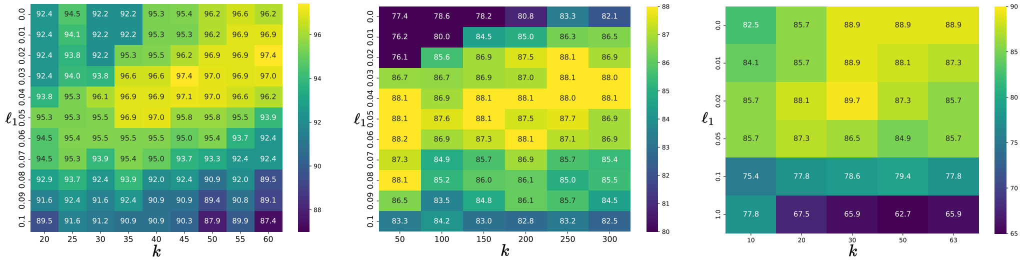
For the UrbanCars dataset, we adhere to the settings described in Li et al. (2023), which involves training a ResNet-50 model pretrained on ImageNet using the SGD optimizer with a learning rate of 10^{-3}, momentum of 0.9, weight decay of 10^{-4}, and a batch size of 128 for 300 epochs. For the Dominoes-CMF dataset, we train a ResNet18 model pretrained on ImageNet for 20 epochs with a batch size of 128 and an SGD optimizer with a learning rate of 10^{-3}, momentum of 0.9, and weight decay of 10^{-4}.

For the image classification tasks, we used random crop and random horizontal flip as data augmentation, similar to Kirichenko et al. (2023). For a fair comparison with the baselines, we did not employ any data augmentation techniques in the process of retraining the last layer of the model.

In EVaLS, we use the implementation of EIIL by `spuco` package (Joshi et al., 2023) for environments inference on the model selection set with 20000 steps, SGD optimizer, and learning rate 10^{-2} for all datasets.

EVaLS and EVaLS-GL For every dataset, EIIL was utilized with a learning rate of 0.01, a total of 20000 steps, and a batch size of 128. The last layer of the model was trained on all datasets using the Adam optimizer. A batch size of 32 and a weight decay of 10^{-4} were used for all datasets. Our method was evaluated on the validation sets of each dataset, considering both fine-tuning and retraining of the last layer. For all datasets, with the exception of MultiNLI and Urbancars, retraining provided superior validation results. The hyperparameter search was conducted over a learning rate range of $[10^{-4}, 10^{-3}]$ and an ℓ_1-regularization coefficient (λ) range of $[0.001, 3]$. For the number of selected samples (k), we ensured that $k \ll |D^{\mathrm{LL}}|$, typically selecting less than 25% of the last-layer split. Notably, as $k \to |D^{\mathrm{LL}}|$, the selected subset inherits the imbalance of the original split, undermining our loss-based balancing strategy. Thus, the optimal k remains significantly smaller than the full split to maintain the effectiveness of our approach.

G.5 Sensitivity to Hyperparameters



(a) Waterbirds (b) CelebA (c) UrbanCars

Figure 8: WGA heatmap on D^{MS} for different hyperparameter settings across various datasets.

The parameters k (the number of selected samples from each loss tail) and λ (the ℓ_1 regularization factor) are automatically selected using the environment/group-based validation scheme proposed in our method. Sensitivity heatmaps in Figure 8 demonstrate the impact of k and λ on the worst-group validation accuracy (WGA) across various datasets. Importantly, our results demonstrate that for most datasets, multiple hyperparameter combinations yield optimal or near-optimal performance, reducing the need for exhaustive searches. This suggests that the hyperparameter tuning process is not prohibitively difficult, and even relatively shallow or targeted hyperparameter searches suffice to identify optimal hyperparameter configurations. The difference in WGA between the best and worst hyperparameter settings for the Waterbirds, CelebA, and UrbanCars datasets is approximately 10%, 16%, and 25%, respectively.

Table 13: Results of DFR and AFR with EIIL-inferred environment for model selection.

Method	Waterbirds	Celeba
DFR (with EIIL)	$\mathbf{92.21 \pm 0.02}$	$\mathbf{85.55 \pm 1.0}$
AFR (with EIIL)	82.6 ± 0.04	72.5 ± 0.01

Table 14: Performance comparison between misclassified sample selection and EVaLS on the Waterbirds, CelebA, and UrbanCars datasets. The mean and standard deviation values are calculated over three runs with different seeds.

Method	Waterbirds		CelebA		UrbanCars	
	Worst	Average	Worst	Average	Worst	Average
Misclassified Selection	$77.8_{\pm 5.2}$	$94.0_{\pm 0.4}$	$85.9_{\pm 1.0}$	$89.4_{\pm 0.8}$	$78.4_{\pm 4.5}$	$86.9_{\pm 1.4}$
EVaLS	$88.4_{\pm 3.1}$	$94.1_{\pm 0.1}$	$85.3_{\pm 0.4}$	$89.4_{\pm 0.5}$	$82.1_{\pm 0.9}$	$88.1_{\pm 0.9}$

H Ablation Study

H.1 Use of EIIL with DFR and AFR

We conducted an ablation study to investigate the impact of using environments inferred from EIIL on model selection. Specifically, we benchmarked the performance of DFR and AFR with EIIL-inferred groups. The results, presented in Table 13, demonstrate the effectiveness of incorporating EIIL-inferred groups in model selection. The results show that while EIIL-inferred groups reduce the performance compared to ground-truth annotations for model selection, they still can be effective for robustness to an extent. Moreover, EVaLS outperforms these two methods when using EIIL inferred environments.

H.2 Comparison of High-Loss and Misclassified-Sample Selection

Several methods, such as JTT (Liu et al., 2021a), rely on misclassified points to address group imbalances by treating these points as belonging to a minority group. To verify the effectiveness of loss-based sampling in comparison with misclassification-based sample selection, we conducted an experiment by replacing loss-based sampling in in EVaLS with selecting misclassified samples and an equal number of randomly chosen correctly classified samples from each class. This results in degraded performance compared to EVaLS on the Waterbirds and UrbanCars datasets, and only a marginal improvement (with higher variance) on CelebA, as summarized in Table 14.

H.3 Other Environment Inference Methods

In addition to EIIL, other environment inference methods could be utilized for partitioning the model selection set into environments.

Error Splitting JTT (Liu et al., 2021a) partitions data into two correctly classified and misclassified sets based on the predictions of a model trained with ERM. We split each of these two sets based on labels of samples, obtaining $|\mathcal{Y}| \times 2$ environments.

Random Linear Classifier on Top of the Features Space uses a random classifier to classify features obtained from a model trained with ERM into correctly classified and misclassified sets. Similar to error splitting, we split the sets based on class labels. The difference between error splitting and random classifier splitting is solely in the reinitialization of the classification layer.

The results for EVaLS-ES (EVaLS+Error Sampling) and EVaLS-RC (EVaLS+Random Classifier) are shown in Table 15. One limitation of error splitting is that in datasets with noisy labels or corrupted images, samples that an ERM model misclassifies may not always belong to minority groups. In these situations, choosing models based on their accuracy on corrupted data could lead to the selection of

Table 15: The performances of three environment inference methods, when combined with loss-based sample selection, are evaluated on spurious correlation benchmarks. The mean and standard deviation values are calculated over three separate runs, each initiated with a different seed.

Method	Waterbirds		CelebA		UrbanCars	
	Worst	Average	Worst	Average	Worst	Average
EVaLS-ES	$82.1_{\pm 1.2}$	$\mathbf{94.3_{\pm 0.04}}$	$48.4_{\pm 11.6}$	$69.5_{\pm 6.5}$	$79.2_{\pm 2.9}$	$86.1_{\pm 0.9}$
EVaLS-RC	$\mathbf{88.7_{\pm 1.0}}$	$94.3_{\pm 1.1}$	$78.1_{\pm 5.1}$	$\mathbf{93.5_{\pm 0.2}}$	$\mathbf{82.4_{\pm 3.2}}$	$\mathbf{88.2_{\pm 0.8}}$
EVaLS	$88.4_{\pm 3.1}$	$94.1_{\pm 0.1}$	$\mathbf{85.3_{\pm 0.4}}$	$89.4_{\pm 0.5}$	82.1 ± 0.9	$88.1_{\pm 0.9}$

models that are not robust to spurious correlations. This is demonstrated by the results of EVaLS-ES on the CelebA dataset.

This shortcoming of error splitting can be alleviated by employing a random classifier instead of the ERM-trained one. Due to the feature-level similarity between minority and majority samples in datasets affected by spurious correlation (Sohoni et al., 2020; Kirichenko et al., 2023; Lee et al., 2023), it is expected that the classifier can differentiate between the groups to some extent. This further supports our claim regarding that the information in the feature space of a trained model could be utilized for achieving robustness against spurious correlations on which a trained model relies. As shown in Table 15, surprisingly, EVaLS-RC produces results that are generally comparable to EVaLS. However, the performance of this method may have high variance, depending on the different initializations of the classifier.

I Computational Resources

Each experiment was conducted on one of the following GPUs: NVIDIA H100 with 80G memory, NVIDIA A100 with 80G memory, NVIDIA Titan RTX with 24G memory, Nvidia GeForce RTX 3090 with 24G memory, and NVIDIA GeForce RTX 3080 Ti with 12G memory.

J Societal Impacts

This paper presents work aimed at advancing the field of Machine Learning by improving a trained model's robustness to spurious correlations it relies on without the need for group annotations. Without such efforts, even if a model becomes robust to a known spurious correlation using current approaches, **a persistent concern remains about the presence of unknown spurious correlations**. **Such correlations may affect the model's predictions and remain undetected, posing significant performance, fairness and safety risks.** While previous methods could not be responsible for achieving robustness to unknown spurious correlations, EVaLS mitigates the effects of known and unknown spurious attributes and has the potential to contribute to more equitable AI systems, particularly in applications like healthcare, hiring, and autonomous systems, where group labels are difficult or costly to obtain due to privacy concerns, logistical challenges, or unknown bias sources. As a post-hoc method, EVaLS is a usable approach for the robustification and improving fairness criteria of models that are already trained and in use in the real world. While our approach offers significant progress in fairness, we acknowledge that models can still inherit biases from data or underspecified training objectives, underscoring the importance of rigorous validation in real-world applications. This work supports broader efforts to develop more trustworthy AI systems, but its application should be carefully considered in context-specific ethical discussions.