Multivariate Group Robustness

Jupinder Parmar [*1] **Vincent Liu** [*1] **Tatsunori Hashimoto** [1]

Abstract

Deep neural networks have been shown to attain high performance on average but poor performance on certain subsets of data, oftentimes because they learn incorrect correlations between the target task and spurious features. Group robustness addresses this by training models to maximize worst-case performance over a set of groups. Prior work in group robustness has only focused on a single spurious correlation per dataset, but this univariate setting is insufficient in cases where models should be robust to many spurious correlations, which can be unknown a priori. To this end, we introduce multivariate group robustness, which presumes the existence of many pairs of spurious correlations simultaneously. We rigorously define and measure the *spuriousness* (strength) of such correlations and verify our procedure by creating a multivariate CelebA dataset for group robustness. We show that this multivariate setup lends itself naturally to multi-task learning; our multi-task baselines show 5-14 percentage point improvements in univariate worst-case accuracy without extra tuning or compute. Finally, we are able to qualify the combinations of spurious correlations that lead to gains in our new multivariate setting, which is an important property for interpretable applications.

1. Introduction

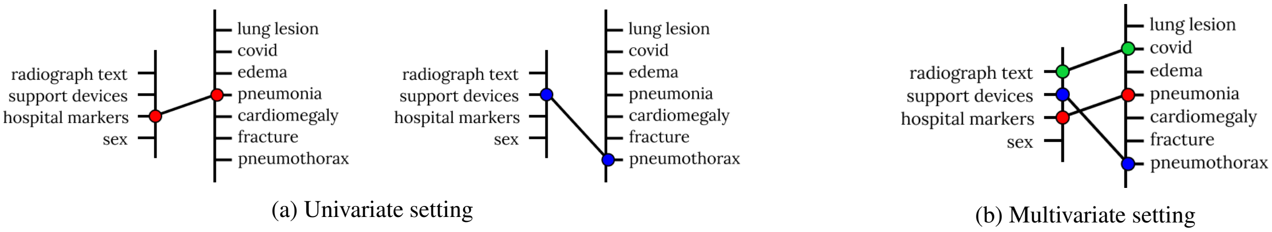
Standard machine learning models are trained with empirical risk minimization (ERM), which minimizes the average training loss to produce models that generalize well to unseen test sets (Hashimoto et al., 2018; Hovy & Søgaard, 2015; Byrd & Lipton, 2019). Despite being highly accurate on average, state-of-the-art models can still incur high error on certain groups of rare and atypical examples. Failures on these groups usually occur when models learn and rely on *spurious correlations*, random associations that hold in the training distribution but not over the true data distribution. For example, in melanoma classification, suspicious lesions are strongly correlated with surgical skin markers (Winkler et al., 2019). A model that learns this spurious correlation would achieve near-perfect average performance, but would fail on groups where this correlation does not hold (i.e. the set of melanoma images without surgical skin markers). Spurious correlations have been found to exist in a variety of domains such as privacy (Leino & Fredrikson, 2020), fairness (Izzo et al., 2020), face recognition (Buolamwini & Gebru, 2018), natural language inference (Gururangan et al., 2018), and image classification (Sagawa et al., 2020a). As neural networks become progressively omnipresent in high-stakes environments such as autonomous vehicles and cancer diagnosis, their robustness to spurious correlations has become a high priority of research.

Though group robustness as a means of addressing spurious correlations is a well-studied field in recent years (Sagawa et al., 2020a; Liu et al., 2021; Goel et al., 2021; Byrd & Lipton, 2019), it has only trained and evaluated models on their robustness to a single spurious correlation per domain. We refer to a spurious correlation between a target task and a spurious attribute in this existing setup as *univariate*. Past work addresses univariate spurious correlations by training models to have low worst-group loss without compromising average loss over a set of non-overlapping groups, usually defined as a Cartesian product between the target task and spurious attribute (Sagawa et al., 2020a). Univariate group robustness methods can be categorized as using group information either in training (Sagawa et al., 2020a; Goel et al., 2021; Zhang et al., 2021), or only in validation for hyperparameter tuning and checkpoint selection (Liu et al., 2021; Sohoni et al., 2020).

However, we argue that the existing framework of univariate group robustness is incomplete as it is constrained to optimize tasks in isolation, when in reality the multivarious interactions of spurious correlation can be leveraged to mitigate the overall influence of individual spurious correlations. For example, in the domain of medical diagnoses from chest x-rays, Oakden-Rayner et al. (2020) found that models trained to diagnose pneumothorax incorrectly rely on the

[*] Equal contribution [1] Department of Computer Science, Stanford University. Correspondence to: Jupinder Parmar <jsparmar@cs.stanford.edu>, Vincent Liu <vliu@cs.stanford.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) Univariate setting

(b) Multivariate setting

Figure 1. Univariate versus multivariate settings. The right axis is an unordered set of potential diagnoses from chest x-rays and the left axis is an unordered set of attributes describing each x-ray. The spurious interactions between attributes and labels can be viewed as a bipartite graph. Current works only address univariate interactions (single-edge subgraphs in Figure 1a), but we propose a setting in which we address multivariate interactions (multi-edge subgraph in Figure 1b) simultaneously over the same distribution $\mathcal{D}$.

presence of chest drains (a device used for the treatment of the condition), while Zech et al. (2019) found that models trained to identify pneumonia rely on hospital-specific features present in the x-rays. In such complex applications, it quickly becomes self-evident that spurious correlations exist in varying degrees of severity and often between attributes that we do not know a priori (hence, spurious). We refer to the concurrence of spurious correlations in the setting above as *multivariate*, existing between many pairs of target tasks and spurious attributes. Since this requires multivariate versions of current datasets, we create a new method for constructing multivariate group robustness datasets and apply it to CelebA (Liu et al., 2015), which has been well-studied in the univariate setting.

To this end, we propose multivariate group robustness, where we seek to mitigate the effect of multiple spurious correlations concurrently as opposed to treating and optimizing for them independently as is done in the univariate case. For example, as shown in Figure 1, instead of diagnosing pneumothorax and pneumonia separately, we assume some underlying relationship in the feature space and optimize them together in a multi-task approach. Intuitively, sharing representation across tasks will regularize the model from depending on task-related spurious correlations – as the sets of spurious features between tasks is likely to be unique – allowing the model to attain better worst-group performance across all tasks. To support multiple spurious correlations, we formulate the optimization problem as multi-task learning (MTL) over some target tasks and their spurious attributes. We show that multivariate models are more group-robust than their univariate counterparts while also making minimal assumptions about the nature of the spurious correlation, which is consistent with other observations of MTL (Caruana, 1997; Ruder, 2017).

In this paper, we describe how to frame the multivariate problem, create a multivariate dataset, and produce new baselines in this new setting. Our contributions are as follows:

- We rigorously define the spuriousness of a correlation, propose a δ metric as a measure of spuriousness, and outline the process of obtaining δ values for each pair of attributes.
- We apply δ to the CelebA dataset and identify 79 pairs of spurious correlations, including the canonical one of hair color and gender as identified in Sagawa et al. (2020a). We verify our method by testing over different pretrained initializations and optimization methods.
- We adapt group robustness baselines from Liu et al. (2021); Idrissi et al. (2022) to optimize multivariate spurious correlations and show improvements in worst-group test set accuracy by 5-14 percentage points.

2. Related Work

In this paper, we focus on group robustness (i.e. learning models that perform equally well across a set of predefined groups in the data) as opposed to other avenues of robustness research such as domain generalization (Muandet et al., 2013; Sun et al., 2020; Zhou et al., 2020). Below, we discuss the different flavors of group robustness research as well as some background on multi-task learning.

2.1. Robustness with group information

Several approaches leverage group labels during training, either to combat spurious correlations or handle shifts in group proportions between train and test distributions. For example, Mohri et al. (2019); Sagawa et al. (2020a); Zhang et al. (2021) minimize the worst-group loss during training, Goel et al. (2021) synthetically expand the minority groups via generative modeling, Shimodaira (2000); Byrd & Lipton (2019); Sagawa et al. (2020b); Idrissi et al. (2022) reweight or subsample to artificially balance the majority and minority groups, and Cao et al. (2019; 2021) impose heavy Lipschitz regularization around minority points. These approaches substantially reduce worst-group error but require group annotations for the entire training set, which can be prohibitively expensive.

2.2. Robustness without group information

We prefer the setting where group annotations are unavailable for the training data and only available on a much smaller validation set. Many approaches for this setting fall under the general distributionally robust optimization (DRO) framework where models are trained to minimize the worst-case loss across all distributions in a ball around the empirical distribution (Duchi et al., 2016; Namkoong & Duchi, 2017; Oren et al., 2019). Pezeshki et al. (2021) modify the dynamics of stochastic gradient descent to avoid learning spurious correlations, Sohoni et al. (2020) automatically identify groups based by clustering the data points, Kim et al. (2019) propose an auditing scheme that searches for high-loss groups defined by a function within a prespecified complexity class and postprocess the model to minimize discrepancies identified by the auditor, and Khani et al. (2019) minimize the variance in the loss across all data points to encourage lower discrepancy in the losses across all possible groups. Another approach is to directly learn how to reweight the training examples either using small amounts of metadata (Shu et al., 2019) or automatically via meta-learning (Ren et al., 2018). Alternatively, prior to training, examples can be either reweighted by class label (Idrissi et al., 2022) or upweighted by misclassified examples from a trained classifier (Liu et al., 2021; Nam et al., 2020). Similarly, Levy et al. (2020) reweight examples, but dynamically in training based on highest loss.

2.3. Multi-task learning

MTL is a simple framework that has shown consistent benefits across domains, tasks, and time (Caruana, 1997; Ruder, 2017). It can improve performance for a main task by incorporating other auxiliary tasks into the loss function (Devlin et al., 2018) or for all tasks (Zhao et al., 2019). We name just a few adaptations of MTL: Myronenko (2018) shows the benefit of an auxiliary reconstruction task on the main segmentation task, Donahue et al. (2021) show how auxiliary multi-resolution spectral prediction tasks improve fidelity of synthetic speech, and Raffel et al. (2022) show how pretraining on multiple text-to-text tasks achieves leading results on downstream benchmarks. Despite immediate gains from the multi-task setting, the specific combinations of tasks can significantly affect performance, making it nontrivial to determine which tasks should be trained together (Fifty et al., 2021; Kumar & Daume III, 2012). Furthermore, MTL can be sensitive to how tasks are weighted: Kendall et al. (2018) show that task weighting by loss as a measure of uncertainty balances training dynamics well; Liu et al. (2019) introduce loss-balanced task weighting, where the per-task weight is set every batch as the task's loss normalized by its loss at the start of the epoch. Task weighting is still an open research problem and many solutions are engineered ad-hoc to address the practical application at hand.

3. Problem Setup

3.1. Spurious correlations

Though spurious correlations have been well-studied in literature, they still have yet to be formally defined. Sagawa et al. (2020b) define them as "misleading heuristics that work for most training examples but do not always hold" and Idrissi et al. (2022) as "patterns that discriminate classes only between specific groups." Hence, we propose the following definition.

Definition 3.1. Two attributes y and a are spuriously correlated with respect to some model f_θ if there exist two distributions $p_{\hat{\mathcal{D}}}(y, a)$ and $p_{\mathcal{D}}(y, a)$ such that

- $p_{\hat{\mathcal{D}}}(y) = p_{\hat{\mathcal{D}}}(a)$ and $p_{\mathcal{D}}(y) = p_{\mathcal{D}}(a)$ are both uniform.
- f_θ attains a much higher loss on $\mathcal{D}$ than it does on $\hat{\mathcal{D}}$.

This definition captures the two main properties of spurious correlations: *correlation* and *severity*. Intuitively, if y and a are spuriously correlated, then $p_{\hat{\mathcal{D}}}$ and $p_{\mathcal{D}}$ differ only in their correlation structure for y and a (the marginals are fixed), but the model relies on this and incurs high (severe) losses on $\mathcal{D}$. We refer to the strength of change in correlation structure of some (y, a) pair as its *spuriousness*.

3.2. Univariate group robustness

We now introduce the univariate problem setup. We seek to classify an input $\mathbf{x} \in \mathcal{X}$ as a binary label $y \in \mathcal{Y}$.[1] We are given a dataset of N points $\{(\mathbf{x}^{(i)}, y^{(i)}, g^{(i)})\}_{i=1}^{N} \sim \mathcal{D}$ sampled from a true distribution. The goal is to learn a classifier, $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$, that incurs low worst-group error over a set of non-overlapping groups $\mathcal{G}$, defined as

$$\min_{\theta} \max_{g' \in \mathcal{G}} \mathbb{E}_{\mathbf{x},y,g \sim \mathcal{D}}[\ell_{0-1}(\mathbf{x}, y; \theta) \mid g' = g] \tag{1}$$

where each example $(\mathbf{x}, y)$ belongs to its group $g \in \mathcal{G}$ and $\ell_{0-1}(\mathbf{x}, y; \theta) = \mathbf{1}[f_\theta(\mathbf{x}) \neq y]$ is the 0-1 loss as in Liu et al. (2021). In our work, each group $g = (y, a)$ is defined by the label y and a spuriously correlated attribute $a \in \mathcal{A}$ as exemplified in Figure 2.

3.3. Multivariate setting

In the multivariate setting, our dataset now has multiple labels $\mathbf{y} = (y_1, ..., y_T)$ such that $\mathbf{y} \in \mathcal{Y}^T$. For each multi-label, we consider some set of attributes, $\mathbf{a} = (a_1, ..., a_T)$, such that each a_j is spuriously correlated to label y_j. We then define a group $g_j = (y_j, a_j)$ as in the univariate setting, such that each multi-label has their associated set of predefined groups $\mathbf{g} = (g_1, ..., g_T)$ where $\mathbf{g} \in \mathcal{G}^T$ and $g_j \in \mathcal{G}_j$.

[1] We work with binary labels, since multiclass labels can be decomposed into a set of binary labels.

Note that $\mathbf{y}$, $\mathbf{a}$, and $\mathbf{g}$ are unordered, but we introduce indexing for convenience. We refer to each distinct unordered $\mathbf{g}$ as a *grouping*.

We essentially create multivariateness by stacking the univariate labels (and their corresponding groups) into a higher dimension. Concretely, in the instance of the well-studied CelebA dataset (Liu et al., 2015), instead of training a model to just predict `Blond Hair` to be robust to the `Male` attribute (Sagawa et al., 2020a; Liu et al., 2021; Idrissi et al., 2022), the model is trained to also jointly predict `Big Lips` to be robust to the `Chubby` attribute. We provide more multivariate examples in Figure 4.

Intuitively, multivariateness prevents the model from overfitting to a specific set of spurious correlations. In the example above, learning the spurious `Male` attribute to predict `Blond Hair` will lead to poor performance on predicting `Big Lips`, and vice versa (learning the spurious `Chubby` attribute to predict `Big Lips` will lead to poor performance on predicting `Blond Hair`), so a model trained to predict both target labels can rely neither on `Male` nor `Chubby` attributes to minimize loss. In reality, relationships between tasks and spurious attributes are probably complex and latent, so the multivariate setting allows us to examine the effect of these interactions.

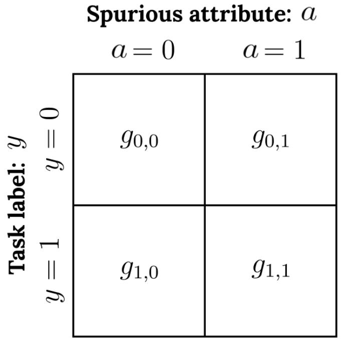
By this construction, we now have a multivariate dataset of N points $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{g}^{(i)})\}_{i=1}^{N} \sim \mathcal{D}$ sampled from a true data distribution. Our goal is to learn a multi-task classifier $f_{\theta,\phi} : \mathcal{X} \mapsto \mathcal{Y}^T$ parameterized by shared θ and task-specific ϕ_j, which achieves low mean worst-group error across all groups simultaneously. We express the optimization as

$$\min_{\theta,\phi} \sum_{j=1}^{T} \max_{g' \in \mathcal{G}_j} \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{g}\sim\mathcal{D}}[\ell_{0-1,j}(\mathbf{x},\mathbf{y};\theta,\phi_j) \mid g' = \mathbf{g}_j] \tag{2}$$

where each point $(\mathbf{x}, \mathbf{y}_j)$ belongs to its group $\mathbf{g}_j \in \mathcal{G}_j$ and $\ell_{0-1,j}(\mathbf{x},\mathbf{y};\theta,\phi) = \mathbf{1}[f_{\theta,\phi}(\mathbf{x})_j \neq \mathbf{y}_j]$.

3.4. Multi-task learning

To accommodate multiple spurious correlations, we frame the multivariate setting as multi-task learning (MTL) over the target labels. Intuitively, by learning a joint representation that is predictive across multiple tasks, the model will be less susceptible to spurious features for each task and learn robust features that are useful for other tasks. As such, MTL acts as a feature regularizer that makes the model robust to spurious attributes without necessarily knowing these attributes a priori. Previous works have shown that features learned by MTL are more robust and indicative of the tasks at hand, which has lead to improvements in domain generalization (Qi et al., 2022; Ruder, 2017; Ghifary et al., 2015).



Task label: y	Spurious attribute: a $a = 0$	$a = 1$
$y = 0$	$g_{0,0}$	$g_{0,1}$
$y = 1$	$g_{1,0}$	$g_{1,1}$

Figure 2. Grid of groups created from the Cartesian product of two binary attributes. In the univariate setting on the CelebA dataset, `Blond Hair` is the task label, which is spuriously correlated with the `Male` attribute.

At its core, a multi-task network minimizes the expected loss across all tasks by using a shared feature backbone which feeds into individual per-task linear heads. Thus, the training objective function becomes:

$$\mathcal{L}_{MTL} = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\hat{\mathcal{D}}}\left[\sum_{j=1}^{T} w_j \ell(\mathbf{x},\mathbf{y}_j;\theta,\phi_j)\right] \tag{3}$$

where f_θ is a shared backbone network, ϕ_j is a linear projection head for task j, ℓ is the loss function, and $w_j > 0$ is the positive weight associated with task j. Since T and w_j are hyperparameters that are still open research questions, we explore optimal settings for both in our experiments.

4. Identifying Spurious Correlations

We now describe how to identify new spurious correlations in a multi-label dataset. Consider two distinct attributes, y and a. We assume these to be binary labels (as is the case in CelebA, where the multi-class label for “hair color” is factored into binary labels for `Blond Hair`, `Brown Hair`, `Black Hair`, `Gray Hair`, etc.) such that their Cartesian product creates 4 groups as shown in Figure 2: $g_{0,0} = (y = 0, a = 0)$, $g_{0,1} = (y = 0, a = 1)$, $g_{1,0} = (y = 1, a = 0)$, and $g_{1,1} = (y = 1, a = 1)$. Then, each sample $(\mathbf{x}, y)$ belongs to a group $g \in y \times a$ and we can measure the expected validation accuracy of the classifier over group g' as

$$\gamma(g';\theta) = \mathbb{E}_{\mathbf{x},y,g\sim\mathcal{D}}[1 - \ell_{0-1}(\mathbf{x},y;\theta) \mid g' = g] \tag{4}$$

4.1. δ metric

In order to identify spurious correlations, we first need to quantify them. We propose the following definition of δ to measure the spuriousness between attributes (y, a).

Definition 4.1. To measure the spuriousness between attributes y and a, we define the δ metric as

$$\delta_{y,a} = |\gamma(g_{0,0}) + \gamma(g_{1,1}) - \gamma(g_{0,1}) - \gamma(g_{1,0})| \quad (5)$$

We claim that δ is an adequate metric for spuriousness by showing that it captures the maximum (worst-case) change in correlation structure induced by moving from $\hat{\mathcal{D}}$ to $\mathcal{D}$. We can describe the correlation between y and a as the optimal transport of loss with marginal constraints. By Definition 3.1, we know that high loss on $\mathcal{D}$ with respect to $\hat{\mathcal{D}}$ indicates a change in correlation structure between y and a as we move from $p_{\hat{\mathcal{D}}}$ to $p_{\mathcal{D}}$. Therefore, we must show that δ measures spuriousness as the change in optimal transport of loss between the two distributions.

Theorem 4.2. *For binary attributes (y, a) on train and test distributions $p_{\hat{\mathcal{D}}}$ and $p_{\mathcal{D}}$, respectively, the maximum change in optimal transport of loss, δ, can be quantified as the difference in group performance between the on- and cross-diagonals.o*

Proof. On distribution D, the optimal transport problem subject to marginal constraints $p_D(y) = p_D(a)$ with cost matrix $\mathbf{C}$ can be expressed as

$$\mathbf{T}_D^* = \min_{\mathbf{T}_D} \langle \mathbf{T}_D, \mathbf{C} \rangle \quad s.t. \quad \begin{cases} \mathbf{T}_D\mathbf{1} = p_D(y) \\ \mathbf{T}_D^\top\mathbf{1} = p_D(a) \end{cases} \quad (6)$$

To maximize the transport of loss from $\hat{\mathcal{D}}$ to $\mathcal{D}$, we can solve for loss-maximizing $\mathbf{T}_{\mathcal{D}}^*$ and loss-minimizing $\mathbf{T}_{\hat{\mathcal{D}}}^*$.

Monge-Kantorovich duality states that any permutation matrix is an optimal transport matrix (Peyré et al., 2019). Since we are working with binary variables, there are only two distinct permutation matrices. If $\mathbf{T}_{\mathcal{D}}^* \neq \mathbf{T}_{\hat{\mathcal{D}}}^*$, then the difference in optimal transport is the absolute difference of diagonals of the minimizer and maximizer, which is exactly the δ characterization. Otherwise, if $\mathbf{T}_{\mathcal{D}}^* = \mathbf{T}_{\hat{\mathcal{D}}}^*$, then the minimizer and the maximizer are the same and there is no difference in optimal transport, which is trivially $\delta = 0$.

The δ value follows by applying $\mathbf{T}_{\mathcal{D}}^* - \mathbf{T}_{\hat{\mathcal{D}}}^*$ to $\mathbf{C}$ and taking the absolute value. Because the negative linear correlation between loss and accuracy is irrelevant in absolute values, we replace $\mathbf{C}_{i,j}$ with $\gamma(g_{i,j})$ to get the δ metric as the absolute difference of on- and cross-diagonal test-set accuracies. □

We note that δ is a heuristic that attempts to uncover spurious correlations as artifacts of what models learn from training data, agnostic to our human perceptions of spurious correlations. For example, δ identifies (`Blond Hair`, `Male`) as a spurious correlation, which we would agree with since they are not correlated in reality. However, δ also identifies (`Blond Hair`, `Gray Hair`) as a spurious correlation, which we would not agree with since they are mutually exclusive in reality. As a result, we see that δ is not perfect – it is unable to identify causality since can only approximate correlation with neural networks. To correct for these situations, we manually validate that all extracted spurious correlations are between non-causal attributes.

4.2. Correcting for small group sizes

We often observe that small group sizes produce spurious correlations. To account for variance from small sample sizes on the validation set (with the smallest groups containing as little as 1% of all examples), we train multiple models with different random seeds and construct 95% confidence intervals of validation accuracies using the Agresti-Coull method (Agresti & Coull, 1998). We outline how to construct these confidence intervals in Appendix A.

Let the validation accuracy of the classifier on group $g_{i,j}$ be $\rho_{i,j}$. To correct for small group sizes, we construct a confidence interval, $[\gamma(g_{i,j})^-, \gamma(g_{i,j})^+]$, and obtain the variance-adjusted estimate clamped to the interval as

$$\bar{\gamma}(g_{i,j}) = \max(\min(\rho_{i,j}, \gamma(g_{i,j})^+), \gamma(g_{i,j})^-) \quad (7)$$

We then compute $\delta_{y,a}$ by replacing $\gamma(g_{i,j})$ with $\bar{\gamma}(g_{i,j})$, which will mitigate the effect of high-variance accuracy computations. From now on, $\delta_{y,a}$ refers to this adjustment.

4.3. Multivariate CelebA

To better understand how underlying characteristics of spurious correlations may help in the multivariate setting, we use CelebA (Liu et al., 2015), a well-studied multi-label dataset for multivariate group robustness. CelebA contains images of the faces of celebrities along with 40 labeled binary attributes ranging from hair color to face shape. Since image classification models have been shown to learn spurious associations of demographic information (Buolamwini & Gebru, 2018), CelebA's large number of labeled attributes allows us to examine multivariate spurious phenomena in depth. Additionally, the spurious correlation (`Blond Hair`, `Male`) has been studied extensively in the univariate setting (Sagawa et al., 2020a; Liu et al., 2021; Idrissi et al., 2022; Sohoni et al., 2020).

To compute δ for each pair of attributes, we fine-tune an ImageNet-pretrained classifier with ERM to classify each individual target attribute, then evaluate each model on the groups created with respect to each of the other 39 attributes ($40 \times 39 = 1560$ values total). Each model is the ResNet-50 architecture (He et al., 2015) used in Sagawa et al. (2020a); Liu et al. (2021), trained with Adam optimizer for 25 epochs with a learning rate of 1e-4 and weight decay of 1e-1.

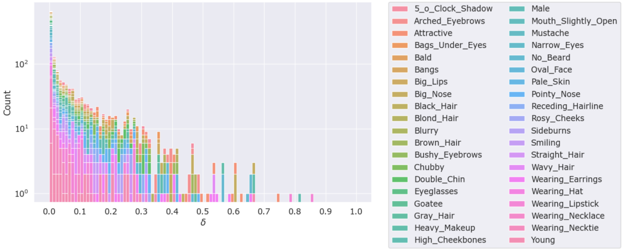


Figure 3. Distribution of δ values with ERM for all 1560 attribute pairs in CelebA. We set $\epsilon = 0.33$ to where we notice a dropoff, but this threshold can be adjusted based on subjective tolerance or application-specific sensitivity to spurious correlations.

Figure 3 shows the distribution of δ values on CelebA. The vast majority of attribute pairs exhibit small δ values, but at around $\delta = 0.33$ we witness the long tail of larger δ pairs. We take this gap to separate spurious from non-spurious correlations, and set $\epsilon = 0.33$ such that all pairs where $\delta > \epsilon$ are spurious correlations. This produces 79 spurious correlations, or around 5% of all possible pairs. Furthermore, we find that $\delta_{\texttt{Blond Hair, Male}} \approx 0.3776$ and confirm past work built on Sagawa et al. (2020a).

We further verify our δ metric by using general pretrained initializations such as CLIP (Radford et al., 2021) and class-balanced optimizations such as SUBY (Idrissi et al., 2022). We show that the spurious correlations identified with δ exist irrespective of the model and its optimization by replicating the above process for a CLIP-pretrained modified ResNet-50 fine-tuned with ERM and an ImageNet-pretrained ResNet-50 fine-tuned with SUBY. Both ablations yield little variation in the distribution of δ values (see Appendix B for details), suggesting that δ measures expected spurious correlations captured by neural networks and is consistent with Section 3.1. We release our new multivariate CelebA dataset with spurious correlation pairs identified from all ablations (ERM, SUBY, and CLIP).

5. Multivariate Group Robustness

The new multivariate CelebA dataset allows us to explore a new set of methods, such as MTL. We want to understand the impact of multivariatness as we move from single-task to multi-task methods. To this end, we extend the univariate baselines from Liu et al. (2021); Idrissi et al. (2022), which belong to the same family of re-weighted sampling in training set without using group labels.

5.1. Multi-task Just Train Twice

In Just Train Twice (JTT), Liu et al. (2021) train an initial classifier f_{id} to generate an error set of misclassified training examples, $\mathcal{E}$. Each example in this error set is upsampled by a factor of λ in training the final classifier f_θ. Intuitively, this upweights examples from groups on which ERM models perform poorly, which serves as a heuristic for identifying examples that contain spurious features.

This, however, is not immediately applicable in the multivariate setting where f_{id} performs T classifications of each input and produces an error set $\mathcal{E}_j$ for each task $\mathbf{y}_j$. As a result, each example is ascribed multiple weights, which need to be mapped to a single weight that is representative proportionally across all tasks. To address the question of how misclassified examples and spurious correlations are related across tasks, we propose the inverse weighting scheme

$$\lambda_i = \begin{cases} \frac{\lambda}{\sum_j \mathbf{1}[\mathbf{x}^{(i)} \in \mathcal{E}_j]} & \mathbf{x}^{(i)} \in \bigcup_j \mathcal{E}_j \\ 1 & \text{else} \end{cases} \tag{8}$$

such that each example $\mathbf{x}^{(i)}$ is upweighted inversely to of the number times it is misclassified. For examples that are never misclassified, we follow Liu et al. (2021) and assign a default weight of $\lambda_i = 1$.

Intuitively, we hypothesize that a frequently misclassified example may be more difficult rather than spurious, so we decrease its importance in training. By employing this weighting scheme in our multi-task version of JTT, we verify whether MTL will still improve worst-group performance despite potentially spurious examples being less important. This would imply that shared representation learning provides benefit where re-weighting based on indicators of spuriousness cannot.

5.2. Multi-task Simple Data Rebalancing

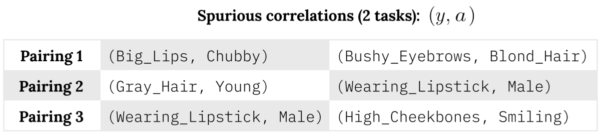
Idrissi et al. (2022) find that balancing training data by classes or groups improves group robustness, and propose downweighting or subsampling majority subsets as simple baselines. We extend class-based subsampling (SUBY) and reweighting (RWY) to the multivariate setting. In the univariate setting, both methods assign a weight to each example based on the proportion of its class in the training set to dictate the example's relative importance in training.

For each $(\mathbf{x}^{(i)}, y^{(i)})$ from training set $\hat{\mathcal{D}}$, the assigned weight is:

$$w_i = \frac{N}{\sum_k^N \mathbf{1}[y^{(k)} = y^{(i)}]} \tag{9}$$

where $N = |\hat{\mathcal{D}}|$. In SUBY, $p_i = 1 - 1/w_i$ is the probability of keeping an example such that the subsampled dataset is class-balanced in expectation. In RWY, w_i is the relative weight used to sample an example such that each batch class-balanced in expectation.

These methods are not directly applicable in MTL as each example will be assigned T weights, one for each task. Data rebalancing in the multivariate setting requires mapping these T weights to a single w , which can then be used



Spurious correlations (2 tasks): (y, a)

Pairing 1	(Big_Lips, Chubby)	(Bushy_Eyebrows, Blond_Hair)
Pairing 2	(Gray_Hair, Young)	(Wearing_Lipstick, Male)
Pairing 3	(Wearing_Lipstick, Male)	(High_Cheekbones, Smiling)

Figure 4. The groupings used in our multi-task baseline experiments ($T = 2$; no common spurious attributes between each pair).

to reweight and subsample the training data in the same manner as in the univariate setting, such that each class for every task is equally represented in the training set. Since $\mathbf{y}$ are multi-dimensional, the per-task weight will be different based on the distribution of class values per task $\mathbf{y}_j$, so there is no closed form mapping to a single $\mathbf{w}$. Instead, we approximate this as a constrained entropy minimization problem:

$$\mathbf{w} = \min_{w} \sum_{i} -w_i \log(w_i) \quad s.t. \quad \begin{cases} \mathbf{Y}^\top \mathbf{w} = c \\ |\mathbf{w}| = 1 \\ \min \mathbf{w} \geq 0 \end{cases} \tag{10}$$

where $\mathbf{Y}$ is the matrix of all multi-labels in the training set and $c = \mathbf{1}/2$ since we want equal representation of binary classes. The entropy minimization formulation encourages the probability distribution $\mathbf{w}$ to be flat and the constraint $\mathbf{Y}^\top \mathbf{w} = c$ ensures that that classes are balanced in expectation after reweighting or subsampling.

6. Experiments

With both multivariate dataset and methods in hand, we explore how working in the multivariate setting can provide gain over the univariate one. Significant gain would verify that the shared representation of multi-task setups improves robustness to spurious features, and would show how MTL can be a straightforward mechanism to improve existing group robustness without any additional tuning or compute.

In all experiments, we assume that group labels are only available on a validation set for hyperparameter tuning and checkpoint selection. We use the ResNet-50 architecture (He et al., 2015), train over 3 seeds, and report 95% confidence intervals as detailed in Appendix A. We only tune hyperpameters for both univariate and multivariate ERM. For RWY, SUBY, JTT we use the univariate hyperparameters identified for CelebA in (Idrissi et al., 2022; Liu et al., 2021) for both univariate and multivariate experiments. We experiment with three forms of task weighting, but find that equal weights across tasks works best. We report all multivariate results with this configuration. Lastly, we only use ERM for all ablative studies. See Appendix C for task weighting details and Appendix D for all hyperparameters.

Table 1. 95% confidence intervals of worst-group accuracy across tasks for the three groupings in Figure 4. Absolute gain is the improvement in accuracy from the univariate to multivariate setting.

Method	Worst-Group Accuracy		Absolute Gain
	Multi	Uni	
ERM	**46.77** ± 3.25	38.82 ± 2.89	+7.95
RWY	**59.01** ± 1.45	53.0 ± 1.87	+6.01
SUBY	**60.37** ± 1.45	55.97 ± 1.92	+4.40
JTT	**61.09** ± 1.43	46.87 ± 0.58	+14.22

6.1. Multivariateness improves over univariateness

We want to understand the immediate gains of the multivariate setting from the univariate one. For simplicity, we explore the least restrictive multivariate scenario of $T = 2$ total tasks. The base case allows us to chalk up all changes in worst-group performance to multivariateness. The 79 spurious correlations identified in Section 4.3 contain 22 unique tasks. We randomly sample a total of three groupings, each containing two spurious correlations (see Figure 4). This results in six unique target labels (a quarter of the 22 unique tasks with spurious correlations), which gives us certainty that these results should hold on the larger population.

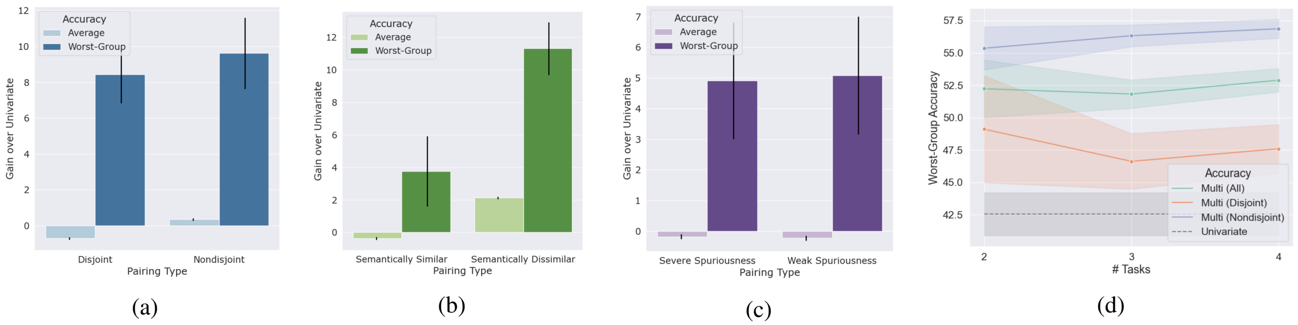
We provide results across our 4 baselines in Table 1 and report the mean over all six spurious correlations in this study. The multivariate baselines consistently outperform the univariate ones and we see absolute gains of up to 14% in worst-group accuracy without compromising average accuracy (Table 2 in Appendix E). Our results indicate that multivariate methods improve performance without any additional tuning, and that a simple way of improving group robustness is by adding more tasks to the training objective.

6.2. Multivariate setting ablations

Since task selection is an open research question in MTL, we also want to examine what combinations of spurious correlation pairs lead to the largest gains. We examine the following scenarios: tasks that have disjoint spurious attributes, tasks that are semantically similar, and tasks with high degrees of spuriousness. For each setting, we randomly sample five groupings that each contain two spurious correlation pairs. We also ablate the number of joint tasks to observe marginal returns of MTL as a function of T.

6.2.1. Tasks with disjoint spurious attributes

Tasks with disjoint spurious attributes have no spurious attributes in common (i.e. attributes are either spuriously correlated to Big Lips or Blond Hair, but not both). We initially hypothesized that the joint representation in the multivariate setting provides benefit due to the fact that the



(a)

(b)

(c)

(d)

Figure 5. Results of multivariate ablations. Figure 5a shows that the relation of spurious attributes across tasks does not significantly affect performance. Figure 5b shows that learning semantically dissimilar tasks together is beneficial. Figure 5c shows that spuriousness is not important to performance. Figure 5d shows additional tasks only benefit in setting of tasks with nondisjoint spurious attributes.

sets of spurious features between tasks are likely to be distinct. Hence, models trained on tasks with disjoint spurious attributes should perform better since there is no common spurious attribute that the model can rely on to predict both tasks well. Figure 5 shows this to be insignificant, as the disjointness of spurious correlations has minimal effect on performance gain. This suggests that the simple addition of another task is enough to improve group robustness.

6.2.2. Semantically dissimilar tasks

We refer to semantically similar tasks as labels that refer to similar attributes of the face (i.e. `Blond Hair` and `Bangs`) and semantically dissimilar tasks as labels which refer to different attributes of the face (i.e. `Blond Hair` and `Big Lips`). Intuitively, semantically dissimilar tasks will require learning more general features which discourages the model from relying on common spurious attributes. Figure 5 confirms this and illustrates that task semantics can lead to an 8% difference in accuracy improvements.

6.2.3. Tasks with severe spuriousness

We define tasks with *severe spuriousness* as spurious correlations whose δ values are above the median (and weak spuriousness as below the median). Though we suspect that tasks with severe spuriousness may be difficult to address due to strong dependencies on their spuriously correlated attributes, we find in Figure 5 that this does not lead to markedly different worst-group accuracy gains. We also see that, in comparison to the other scenarios, task spuriousness is not as of an important criterion in multivariateness.

6.2.4. Number of tasks

Since MTL scales with the number of target labels available, it is insightful to see when MTL stops providing marginal performance gain. Our goal is to understand the impact of increasing the number of tasks, and so we construct ablations on T by sequentially adding tasks to an initial set of 2 tasks so that we can evaluate our models on the same tasks for each experiment. See Appendix E for details.

We expect that as we add tasks, the interaction between spurious attributes plays a bigger role in overall robustness, so we ablate T under different relationships between spurious attributes. Figure 5d highlights the change in worst-group accuracy across our base $T = 2$ tasks as we increase T. On average, there is no benefit from larger T, but tasks with nondisjoint spurious attributes show positive marginal gain, while tasks with disjoint spurious attributes show negative marginal gain. This indicates that if we choose tasks at random for a grouping without full understanding of the interactions between their spurious correlations then the number of tasks is not an important aspect of performance.

7. Conclusion

In this paper, we introduce multivariate group robustness, a generalization of existing univariate work to address multiple target labels with their own associated spurious attributes. We propose the first formal definition and metric for spurious correlations. We connect our method to definition via rigorous theoretic foundation and identify 79 new spurious correlations in our new multivariate CelebA dataset. We demonstrate how to extend univariate baselines to the multivariate setting and show that this universally leads to significant gains in worst-group accuracy with no degradation of average accuracy. Our ablation experiments also give insight into the combinations of tasks and spurious attributes that lead to the largest gains in the multivarate setting.

In conclusion, we demonstrate the importance of multivariate group robustness and how it unifies the approaches in past works. Though some multivariate assumptions can be limiting, such as access to many high-quality labels, we cite the universality of multi-task learning to emphasize the need for future work here. We release one multivariate dataset and hope that others will expand our formalizations to create additional benchmarks.

References

Agresti, A. and Coull, B. A. Approximate is better than ”exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998. ISSN 00031305. URL http://www.jstor.org/stable/2685469.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018.

Byrd, J. and Lipton, Z. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pp. 872–881. PMLR, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Cao, K., Chen, Y., Lu, J., Arechiga, N., Gaidon, A., and Ma, T. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2021.

Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Donahue, J., Dieleman, S., Binkowski, M., Elsen, E., and Simonyan, K. End-to-end adversarial text-to-speech. In *International Conference on Learning Representations*, 2021.

Duchi, J., Glynn, P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Fifty, C., Amid, E., Zhao, Z., Yu, T., Anil, R., and Finn, C. Efficiently identifying task groupings for multi-task learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders, 2015. URL https://arxiv.org/abs/1508.07680.

Goel, K., Gu, A., Li, Y., and Re, C. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Hovy, D. and Søgaard, A. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 483–488, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2079.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *First Conference on Causal Learning and Reasoning*, 2022.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. *arXiv preprint arXiv:2002.10077*, 2020.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Khani, F., Raghunathan, A., and Liang, P. Maximum weighted loss discrepancy. *CoRR*, abs/1906.03518, 2019.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Blackbox post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Kumar, A. and Daume III, H. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

Leino, K. and Fredrikson, M. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC’20, USA, 2020. USENIX Association. ISBN 978-1-939133-17-5.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Liu, S., Liang, Y., and Gitter, A. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *AAAI*, pp. 9977–9978, 2019.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation, 2013.

Myronenko, A. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pp. 311–320. Springer, 2018.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4227–4237, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1432.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Qi, L., Yang, H., Shi, Y., and Geng, X. Multimatch: Multi-task learning for semi-supervised domain generalization, 2022. URL https://arxiv.org/abs/2208.05853.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jun 2022. ISSN 1532-4435.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020a.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020b.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., and Hardt, M. Test-time training for out-of-distribution generalization, 2020.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., and Haenssle, H. A. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10):1135–1141, 10 2019. ISSN 2168-6068. doi: 10.1001/jamadermatol.2019.1735. URL https://doi.org/10.1001/jamadermatol.2019.1735.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med*, 15(11): e1002683, 2019.

Zhang, J., Menon, A. K., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*, 2021.

Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., Kumthekar, A., Sathiamoorthy, M., Yi, X., and Chi, E. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 43–51, 2019.

Zhou, K., Yang, Y., Hospedales, T., and Xiang, T. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pp. 561–578. Springer, 2020.

A. Agresti-Coull confidence intervals

For each seed k, we train a model parameterized by θ_k. Let p_k be the model's accuracy over some set of examples that we would like to evaluate (for accuracy on some group g, we have $p_k = \gamma(g; \theta_k)$) and n be the number of samples in this set. Using Gaussian approximation, we compute the empirical mean $\hat{\mu}_k$ and variance $\hat{\sigma}_k^2$ as

$$\hat{\mu}_k = p_k, \qquad \hat{\sigma}_k^2 = \frac{p_k(1-p_k)}{n} \tag{11}$$

Aggregating over the seeds, we get

$$\hat{\sigma}^2 = \frac{1}{\Sigma_k(1/\hat{\sigma}_k^2)}, \qquad \hat{\mu} = \hat{\sigma}^2 \Sigma_k \left(\hat{\mu}_k/\hat{\sigma}_k^2\right) \tag{12}$$

The confidence interval can then be computed as $\hat{\mu} \pm z\hat{\sigma}$ where z is the $1 - \alpha/2$ percentile of the normal distribution. We report all results over $k = 3$ seeds with $\alpha = 0.05$, which corresponds to 95% confidence intervals.

B. Spurious correlation identification

Each SUBY model was trained for 25 epochs using Adam optimizer with a learning rate of 1e-4, weight decay of 1e-1, and a batch size of 64. Each CLIP model was trained for 50 epochs using Adam optimizer with a learning rate of 1e-4, weight decay of 1e-1, and a batch size of 128. The histogram of recovered δ values for both CLIP and SUBY are shown in Figure 6 and Figure 7 respectively.

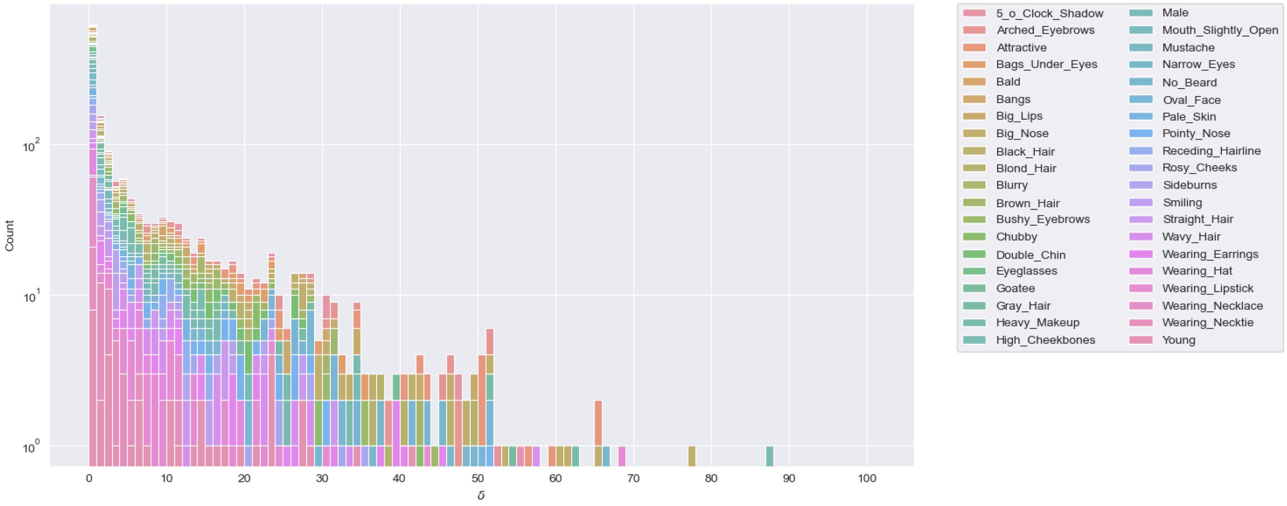


Figure 6. CLIP identified δ values for all 40×39 pairs in CelebA. The resulting distribution, and identified spurious correlation pairs with large δ values, is similar to that identified by ERM.

C. Task weighting schemas

C.1. Equal weighting

The most naive form of task weighting in multi-task learning consists of setting each weight associated to one of the T tasks to an equal value such that $\sum_{i=1}^{T} w_i = 1$. Hence, we get that each $w_i = \frac{1}{T}$

C.2. Weighting based on delta

We additional consider weighting each spurious correlation pair in a grouping by its associated delta value. Hence, if for each of the T tasks we have an associated δ_{y_i,a_i} then the weight associated to task i is defined to be: $w_i = \text{Softmax}(\boldsymbol{\delta})_i$ where $\boldsymbol{\delta} = (\delta_{y_1,a_1}, ..., \delta_{y_T,a_T})$

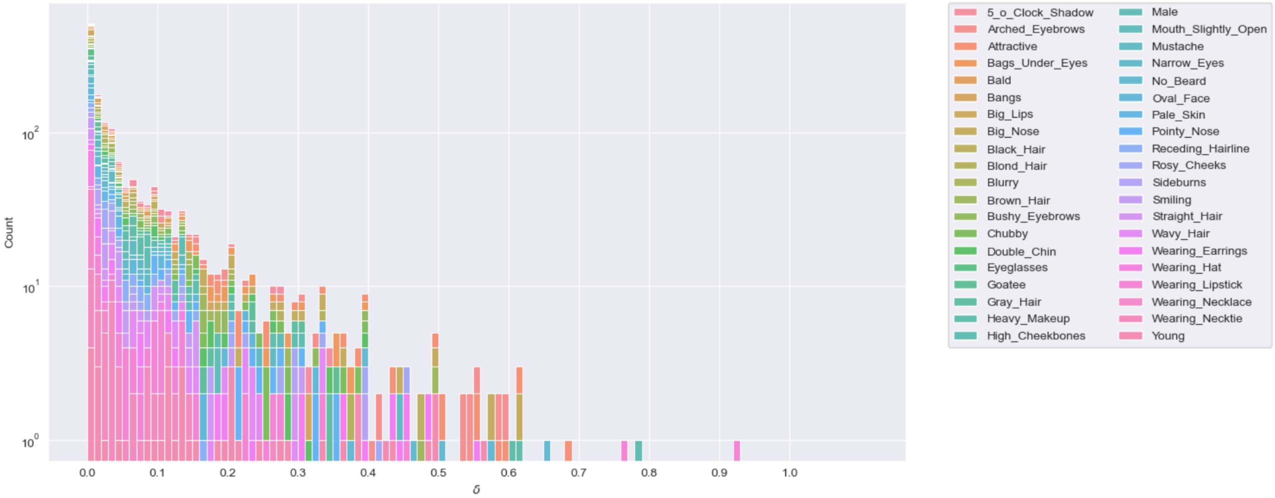


Figure 7. SUBY identified δ values for all 40×39 pairs in CelebA. The resulting distribution, and identified spurious correlation pairs with large δ values, is similar to that identified by ERM.

C.3. Loss balanced task weighting

After each batch in training, loss balanced task weighting (Liu et al., 2019) updates the weights for a each task based on the loss ratio between the loss of the current batch and the loss of the initial batch. This ratio acts as a metric for how well the model has trained for that given task. See Algorithm 1 for details.

Algorithm 1 Loss-Balanced Task Weighting

```
Given $t$ tasks and hyperparameters $\alpha, \eta$
Initialize the model $\theta, \phi \in \Theta$
for each epoch $i$ do
  for each batch $j$ do
    Compute loss $\ell^{(j)} := \ell^{(j)}(x, y; \theta, \phi) \in \mathbb{R}^t$.
    if $j = 0$ then
      Store the first batch loss as $\ell^{(0)} := \ell^{(j)}(x, y; \theta, \phi)$.
    end if
    for each task $k$ do
      Set the task weight $w_k := \left(\frac{\ell_k^{(B)}}{\ell_k^{(0)}}\right)^{\alpha}$
      Update weighted loss $\ell^{(B)} := \sum_k w_k \ell_k^{(B)}$
    end for
    Update parameters $\theta, \phi$ with respect to $\ell^{(B)}$
  end for
end for
```

D. Experimental hyperparmeter values

When tuning ERM in both the univariate and multivariate settings we search over the following values:

- Weight Decay: [1e-4, 1e-3, 1e-2, 1e-1]
- Learning Rate: [1e-5, 1e-4, 1e-3]
- Batch Size: [32, 64, 128]

We found that the best parameters for univariate ERM were a weight decay of 1e-4, a learning rate of 1e-4, a batch size of 128, and training for 50 epochs. The grid search resulted in multivarite ERM using a weight decay of 1e-2, a learning rate of 1e-4, a batch size of 32, and training for 50 epochs.

For all other methods we use the following parameter values:

- Multivariate and Univariate SUBY: Weight Decay=1e-2, Learning Rate=1e-3, Batch Size=128, Epochs=60
- Multivariate and Univariate RWY: Weight Decay=1e-2, Learning Rate=1e-4, Batch Size=2, Epochs=60
- Multivariate and Univariate JTT: Weight Decay=1e-1, Learning Rate=1e-5, Batch Size=128, Epochs=50, λ=1

E. Experimental results and ablations

E.1. Multivariate vs univariate results

Table 2 highlights that the multivariate setting does not compromise average accuracy performance compared to the univariate setting – in some baselines, it even improves performance.

Table 2. 95% confidence intervals of average accuracy across tasks for the three groupings in Figure 4. Absolute gain represents the increase in accuracy by moving from the univariate setting to the multivariate setting.

METHOD	AVERAGE ACCURACY		ABSOLUTE GAIN
	MULTI	UNI	
ERM	87.35± 0.10	**87.79** ± 0.10	-0.44
RWY	**85.08** ± 0.12	84.36 ± 0.11	+0.72
SUBY	83.98 ± 0.12	**85.02**± 0.11	-1.04
JTT	**79.23**± 0.13	70.32± 0.14	+8.89

E.2. Num tasks ablation

We explicitly outline the experimental setting of Section 6.2.4 below:

Letting Multi(n) refer to the multivariate setting where there are n spurious correlation pairs, we generate all groupings as

1. Sample two groupings for Multi(2) from the set of all groupings of size 2.
2. For each of the two Multi(2) groupings, sample three spurious correlations from our set of 79 spurious correlations to form groupings for Multi(3).
3. For each of the six Multi(3) groupings, sample one more spurious correlations to form groupings for Multi(4).
4. For each task used in Multi(2), compare Multi(n) for $n \in [2, 3, 4]$ against the univariate performances.

As trained models may perform at different accuracies on different tasks, we ensure that at each ablation we evaluate our models on the same initial two tasks.