AVOIDING SPURIOUS CORRELATIONS VIA LOGIT CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Empirical studies suggest that machine learning models trained with empirical risk minimization (ERM) often rely on attributes that may be spuriously correlated with the class labels. Such models typically lead to undesired and poor performance during inference for data lacking such correlations and generalize even worse when more training data present spurious correlations. In this work, we explicitly consider the presence of the potential spurious correlations exist in the majority of training data. Unlike existing approaches which use the ERM model outputs to detect the samples without spurious correlations, and heuristically upweight or upsample those samples, we propose the logit correction (LC) loss, a simple yet effective improvement on the softmax cross-entropy loss, to correct the sample logit. We demonstrate that minimizing the LC loss is equivalent to maximizing the group-balanced accuracy, thus the proposed LC could mitigate the negative impacts of spurious correlations in the majority of samples. Our extensive experimental results further reveal that the proposed LC loss outperforms the SoTA solutions on multiple popular benchmarks by a noticeable large margin, an average 5.5% absolute improvement, without access to spurious attribute labels. LC is also competitive with oracle methods that make use of the attribute labels.

1 INTRODUCTION

A robust machine learning model for daily use (e.g., a self-driving car) must be designed to comprehend its surroundings in rare conditions that may not have been well-represented in its training set. However, it is observed that deep neural networks are negatively affected by spurious correlations between observed features and class labels that hold for well-represented groups but not for rare groups. For example, consider classifying stop signs versus other traffic signs in autonomous driving, 99% of the stop signs in the United States are of a octagonal shape in red. A model trained with standard empirical risk minimization (ERM) may learn models with low average training error but relying on the spurious background attribute instead of the desired "STOP" text on the sign, resulting in a high average accuracy but a low worst-group accuracy (e.g. making errors on yellow color or faded stop signs). This demonstrates a fundamental issue that models trained on such datasets could be systematically biased due to spurious correlations presented in their training data (Ben-Tal et al., 2013; Rosenfeld et al., 2018; Beery et al., 2018; Zhang et al., 2019). Mitigating such biases is critical in many fields including but not limited to, algorithmic fairness (Du et al., 2021), machine learning in healthcare (Oakden-Rayner et al., 2020), machine learning in public policy Rodolfa et al. (2021) etc.

Formally, the spurious correlation occurs when the target label is mistakenly associates with one or more confounding factors presented in the training data. The group of samples in which the spurious correlations occur is often called the *majority group* since spurious correlations are expected to occur in most samples, while the *minority groups* contain samples whose features are not spuriously correlated. It is explored that performance degradation of ERM on a dataset with spurious correlation (Nagarajan et al., 2021; Nguyen et al., 2021) are caused by two main reasons: 1) the geometric skew and 2) the statistical skew. The geometric skew results from the fact that the classification margin on the minority group of a robust classifier should be much larger than that of the majority group (Nagarajan et al., 2021). However, a classifier trained with ERM maximizes margins and therefore leads to equal training margins for majority and minority groups. The statistical skew is caused by slow convergence of gradient descent which results in that the network may firstly pick up

the "easy-to-learn" spurious attribute instead of the true label information and rely on it until being trained for an exponentially long time.

Many existing approaches considers the absence of group information, and thus first detect the minority group (Nguyen et al., 2021; Liu et al., 2021b; Nam et al., 2020) and then upweight/upsample the samples in the minority group during training (Li & Vasconcelos, 2019; Nam et al., 2020; Lee et al., 2021; Liu et al., 2021a). While intuitive, upweighting only addresses the statistical skew (Nguyen et al., 2021), and it is often hard to define the weighted loss with an optimal upweighting scale in practice. Following Menon et al. (2013); Collell et al. (2016) on learning from imbalanced data, we argue that the goal of training a debiased model is to achieve a high average accuracy over all groups (Group-Balanced Accuracy, GBA, defined in Sec. 3), implying that the training loss should be Fisher consistent with GBA (Menon et al., 2013; Collell et al., 2016). In other words, the minimizer of the loss function should be the maximizer of GBA.

In this paper, we revisit the logit adjustment method (Menon et al., 2021) for long tailed datasets, and propose a new loss called logit correlation (LC) to reduce the impact of spurious correlations. We show that the proposed LC loss is able to mitigate both the statistical and the geometrical skews that cause the performance degradation. More importantly, under mild conditions, its solution is Fisher consistent for maximizing GBA. In order to calculate the corrected logit, we study the spurious correlation and propose to use the outputs of the ERM model to estimate the group priors. To further reduce the geometrical skew, based on MixUp (Zhang et al., 2018), we propose a simple yet effective method called Group MixUp to synthesize samples from the existing ones and thus increase the number of unique samples in the minority groups.

The main contributions of our work include:

- We propose logit correction loss to mitigate spurious correlations during training. The loss ensures the Fisher consistency with GBA and alleviates statistical and geometric skews.
- We propose the Group MixUp method to increase the diversity of the minority group and further reduce the geometrical skew.
- The proposed method significantly improves GBA and the worst-group accuracy when the group information is unknown. With only 0.5% of the samples from the minority group, the proposed method improves the accuracy by 6.03% and 4.61% on Colored MNIST dataset and Corrupted CIFAR-10 dataset respectively over the state-of-the-arts.

2 Related Work

Spurious correlations are ubiquitous in real-world datasets. A typical mitigating solution requires to first detect the minority groups and then design a learning algorithm to improve the group-balanced accuracy and/or the worst-group accuracy. We review existing approaches based on these two steps.

Detecting Spurious Correlations. Early researches often rely on the predefined spurious correlations (Kim et al., 2019; Sagawa et al., 2019; Li & Vasconcelos, 2019). While effective, annotating the spurious attribute for each training sample is very expensive and sometimes impractical. Solutions do not require the spurious attribute annotation attract a lot of attention nowadays. Many of the existing works (Sohoni et al., 2020; Nam et al., 2020; Liu et al., 2021a; Zhang et al., 2022) assume that the ERM model will be prone to the spurious attribute (but may still learn the core features Kirichenko et al. (2022)), thus "hard" examples (whose predicted labels are conflicting with the ground-truth label) are likely to be in the minority group. Sohoni et al. (2020); Seo et al. (2022), on the other hand, propose to estimate the unknown group information by clustering. Our work follows the path of using the ERM model.

Mitigating Spurious Correlations. Previous works (Nagarajan et al., 2021; Nguyen et al., 2021) show that the geometric skew and the statistical skew are two main reasons hurting the performance of the conventional ERM model. Reweighting (resampling) which assigns higher weights (sampling rates) to minority samples is commonly used to remove the statistical skew (Li & Vasconcelos, 2019; Nam et al., 2020; Lee et al., 2021; Liu et al., 2021a). While intuitive, reweighting has limited effects to remove the geometric skew (Nagarajan et al., 2021). Also, there are surprisingly little discussions on how to set the optimal weights. We argue that the reweighting strategy should satisfy the Fisher consistency (Menon et al., 2013), which requires that the minimizer of the reweighted loss is the

maximizer of the balanced-group accuracy (see Sec. 4.1). On the other hand, synthesizing minority samples/features is widely utilized in removing the geometric skew. Minderer et al. (2020); Kim et al. (2021) propose to directly synthesize minority samples using deep generative models. While synthesizing the raw image is intuitive, the computation complexity can be high. DFA (Lee et al., 2021) mitigates this issue by directly augmenting the minority samples in the feature space.

DFA (Lee et al., 2021) is the most related work to our approach. It applies reweighting to reduce the statistical skew and feature swapping to augment the minority feature thus removing the geometric skew. However, our approach uses logit corrected loss and is proved to be Fisher consistent with the group-balanced accuracy. The proposed logit corrected loss has a firmer statistical grounding and can reduce both the statistical skew and the geometric skew. By cooperating the logit corrected loss and the proposed Group MixUp, our approach outperforms DFA, especially on the dataset containing very few minority samples.

3 Problem Formulation

Let's first consider a regular multi-class classification problem. Given a set of n training input samples $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}, i = 1, \dots, n$, where, $\mathbf{x} \in \mathbb{R}^d$ has a input dimension of d and $y \in \mathcal{Y} = \{1, \dots, L\}$ with a total number of L categories. Our goal is to learn a function (neural network), $f(\cdot) : \mathcal{X} \to \mathbb{R}^L$, to maximize the classification accuracy $P_{\mathbf{x}}(y = \arg\max_{y' \in \mathcal{Y}} f_{y'}(\mathbf{x}))$. With ERM, we typically minimize a surrogate loss, *e.g.*, softmax cross-entropy $\mathcal{L}_{CE}(\cdot)$ where,

$$\mathcal{L}_{CE}(y, f(\mathbf{x})) = \log[\sum_{y'} e^{f_{y'}(\mathbf{x})}] - f_y(\mathbf{x}). \tag{1}$$

We assume there is a spurious attribute $\mathcal{A}$ with K different values in the dataset. Note that K and the number of classes L may not be equal. We define a combination of one label and one attribute value as a group $g = (a, y) \in \mathcal{A} \times \mathcal{Y}$. The spurious correlation means an attribute value a and a label y commonly appears at the same time. Different from ERM, the goal of training a model to avoiding spurious correlation is to maximize the Group-Balanced Accuracy (GBA):

$$GBA(f) = \frac{1}{KL} \sum_{g \in \mathcal{G}} \mathbf{P}_{\mathbf{x}|(y,a)=g}(y = \arg\max_{y' \in \mathcal{Y}} f_{y'}(\mathbf{x})). \tag{2}$$

Note that since the attribute value for each sample is unknown during training, in order to distinguish the label and the spurious attribute, it assumes to have some samples without the spurious correlation.

4 Our approach: Logit Correction

We apply a two-branch network (as shown in Figure 1). The top branch (denoted as $\hat{f}(\cdot)$) is a network trained with ERM using generalized cross-entropy (GCE) loss Zhang & Sabuncu (2018):

$$\mathcal{L}_{GCE} = \frac{1 - \hat{f}(\mathbf{x})^q}{q}, \tag{3}$$

where $\hat{f}(\mathbf{x})$ represents the probability outputs of the ERM model, $q \in [0, 1)$ is a hyperparameter. Compared to the standard cross-entropy loss, the gradient of GCE loss upweights examples where $\hat{f}(x)$ is large, which intentionally biases $\hat{f}$ to perform better on majority (easier) examples and poorly on minority (harder) examples. The second (bottom) branch is trained to learn from the first branch. To be more specific, we use the probability output of the first branch to correct the logit output of the second branch. We further adopt Group MixUp to increase the number of unique examples in the minority groups. The details of the method are demonstrated in the following sections.

4.1 Logit Correction as maximizing GBA

Recall that our goal is to maximize GBA in Eq 2 which depends on the (unknown) underlying distribution $\mathbf{P}(x, y, a)$, the Bayes-optimal prediction function under this setting is $f^* \in \arg\max_f GBA(f)$.

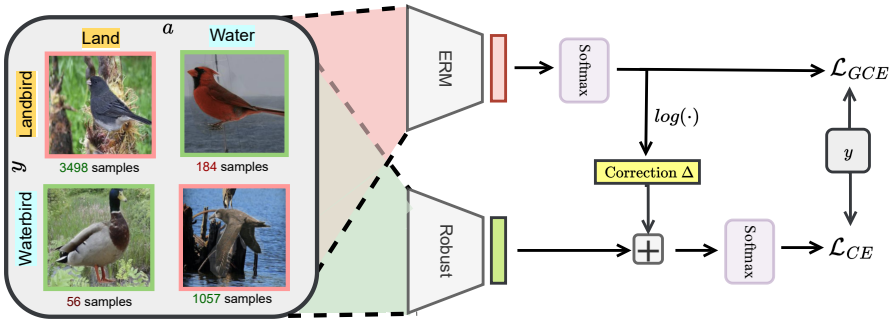


Figure 1: The overview of our proposed logit correction approach on the Waterbirds dataset, where the background (water/land) is spuriously correlated with the foreground (waterbird/landbird). Most training samples belong to the group where the background matches the bird type (highlighted in red); While only a small fraction belong to the groups where the background mismatches the bird type (highlighted in green). The ERM network and the robust network are trained simultaneously. The ERM network is trained with a generalized cross-entropy (GCE) loss to be intentionally biased to the majority group. The logit correction loss corrects the logits of the robust network by a term Δ which is produced by the predictions of the ERM network. The robust network is trained with the standard cross entropy loss after logit correction.

Proposition 1. *Let $P(y, a)$ be the prior of group (y, a), and $P(y, a|\mathbf{x})$ is the true posterior probability of group (y, a) given $\mathbf{x}$, the prediction:*

$$\arg\max_{y\in\mathcal{Y}} f_y^*(\mathbf{x}) = \arg\max_y \sum_a \frac{P(y, a|\mathbf{x})}{P(y, a)} = \arg\max_y \sum_a \frac{P(y|a, \mathbf{x})P(a|\mathbf{x})}{P(y, a)} \tag{4}$$

is the solution to Eq. 2. See proof in Appendix A.

Assume each example $\mathbf{x}$ can only take one spurious attribute value (e.g. waterbirds can either be on water or on land, and can not on both), that is to say, the prior probability $P(a|\mathbf{x})$ is 1 when the spurious attribute $a = a_{\mathbf{x}}$ and 0 otherwise. We have

$$\arg\max_{y\in\mathcal{Y}} f_y^*(\mathbf{x}) = \arg\max_y P(y|a_{\mathbf{x}}, \mathbf{x})/P(y, a_{\mathbf{x}}). \tag{5}$$

Note that although $a_{\mathbf{x}}$ is unknown in the dataset, it can be estimated using the outputs of ERM model (see Sec. 4.2).

Because we are using the second branch to estimate the posterior probability $P(y|a_{\mathbf{x}}, \mathbf{x})$, supposing the underlying class probability $P(y|a_{\mathbf{x}}, \mathbf{x}) \propto \exp(f(\mathbf{x}))$ for an (unknown) scorer f, we can rewrite Eq. 5 as

$$\begin{aligned}\arg\max_y P(y|a, \mathbf{x}) &= \arg\max_{y\in\mathcal{Y}} \exp(f_y(\mathbf{x}))/P(y, a_{\mathbf{x}}),\\ &= \arg\max_{y\in\mathcal{Y}} f_y(\mathbf{x}) - \ln P(y, a_{\mathbf{x}}).\end{aligned} \tag{6}$$

In practice, we could use the corrected logits $f(\mathbf{x}) + \Delta_{y,a_{\mathbf{x}}}$ instead of the original logits $f(\mathbf{x})$ to optimize the network, where $\Delta_{y,a_{\mathbf{x}}}$ are estimates of the group priors $\mathbf{P}(y, a_{\mathbf{x}})$. And then we can predict the label as usual (Menon et al., 2021). The logits corrected softmax cross entropy function can be written as,

$$\mathcal{L}_{LC}(y, f(\mathbf{x})) = \log[\sum_{y'} e^{f_{y'}(\mathbf{x})+\ln\Delta_{y',a_{\mathbf{x}}}}] - (f_y(\mathbf{x}) + \ln\Delta_{y,a_{\mathbf{x}}}). \tag{7}$$

We show that when $P(a|\mathbf{x})$ follows a one-hot categorical distribution, Eq. 7 is Fisher consistent with maximizing the GBA (Eq. 2). Intuitively, the more likely the combination of $(y, a_{\mathbf{x}})$ appears in the training dataset, the less we subtract from the original logits. Meanwhile, the less likely the combination of $(y, a_{\mathbf{x}})$ appears in the training dataset, the more we subtract from the original logits, resulting in less confidence in the prediction. It leads to larger gradients for samples in the minority

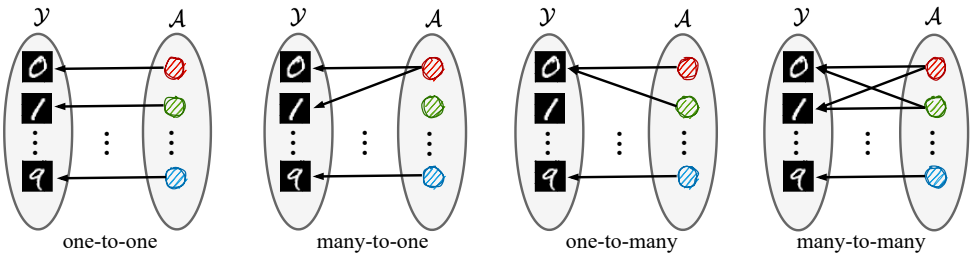


Figure 2: Example of different situations for spurious correlation that can be considered for the colored MNIST dataset. Spurious correlation existed in the majority groups: for one-to-one situation, most samples of one digit has a distinct color; for many-to-one situation, multiple digits are colored by the same color; for one-to-many situation, one digit can be colored by several different colors. Many-to-many situation allows both one-to-many and many-to-one situations. Note that the figure only shows the spurious correlation not all possible combinations.

group, making the network learns more from the minority group. To this end, the logit corrected loss helps reduce the statistical skew. Moreover, Eq. 7 can be further rewritten as

$$\mathcal{L}_{LC}(y, f(\mathbf{x})) = \log\left(1 + \sum_{y' \neq y} e^{f_{y'}(\mathbf{x}) - f_y(\mathbf{x}) + \ln\left(\Delta_{y', a_\mathbf{x}} / \Delta_{y, a_\mathbf{x}}\right)}\right). \tag{8}$$

It's a pairwise margin loss (Menon et al., 2013), which introduces a desired per example margin $\ln\left(\Delta_{y', a_\mathbf{x}} / \Delta_{y, a_\mathbf{x}}\right)$ into the softmax cross-entropy. A minority group example demands a larger margin since the margin is large when $\Delta_{y, a_\mathbf{x}}$ is small. To this end, LC loss is able to mitigate the geometric skew resulting from maximizing margins. We also empirically compare the training margins of minority groups and majority groups in Figure 4 for different methods. In the next section, we will introduce given a sample $\mathbf{x}$ in the training dataset, how to estimate $a_\mathbf{x}$ and the group prior.

4.2 Estimating the Group Prior

We first analyze different spurious correlation relations. The spurious correlation between the target label and the spurious attribute can be categorized into 4 different types (Figure 2): 1) one-to-one, where each label only correlates with one attribute value and vice versa; 2) many-to-one, where each label correlates with one attribute value, but multiple labels can correlate with the same attribute value; 3) one-to-many, where one label correlates with multiple attribute values; and 4) many-to-many, where multiple labels and multiple attribute values can correlate with each other. In the next section we will first discuss the most common one-to-one situation and then extend it to other situations.

4.2.1 One-to-one

One-to-one is the most common situation studied in the previous works, *e.g.*, the original Colored MNIST dataset. To estimate the group prior probability $P(y, a)$, we consider

$$P(y, a) = \int_\mathbf{x} P(y, a, \mathbf{x}) d\mathbf{x} = \int_\mathbf{x} P(y, a|\mathbf{x}) P(\mathbf{x}) d\mathbf{x} \approx \frac{1}{N} \sum_\mathbf{x} P(y, a|\mathbf{x}). \tag{9}$$

The last approximation is to use the empirical probability estimation to estimate the prior, where N is the number of total samples. It's impractical to use the whole dataset to estimate the group prior after each training iteration. In practice, we try different estimation strategies and find out that using a moving average of the group prior estimated within each training batch seems to produce reasonable performance (see Figure 3). We apply this method to all the experiments in the paper.

For a training sample of $(\mathbf{x}_i, y_i, a_i)$, since the label y_i is known, we have

$$P(y, a|\mathbf{x}_i) = \begin{cases} P(a|\mathbf{x}_i), & \text{if } y = y_i. \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Since the spurious correlation is one-on-one, the number of category and the number of different attribute values are the same. Without loss of generality, we assume that the j-th category ($y^{(j)}$) is

correlated with the j-th attribute value ($a^{(j)}$), where $j = [1, \dots, L]$ and $L = K$. L and K are the number of categories and the number of different attribute values respectively.

Since the ERM network would be biased to the spurious attribute instead of the target label, the prediction of the ERM network can be viewed as an estimation of $P(a|\mathbf{x}_i)$. Formally, denote the output logits of the ERM network as $\hat{f}(\mathbf{x}_i)$ and the j-th element of $\hat{f}(\mathbf{x})$ is denoted as $\hat{f}_j(\mathbf{x}_i)$, we have

$$P(a = a^{(j)}|\mathbf{x}_i) = \frac{\exp(\hat{f}_j(\mathbf{x}_i))}{\sum_{k=1}^{K} \exp(\hat{f}_k(\mathbf{x}_i))} \tag{11}$$

Given Eq. 9 to Eq. 11, we can estimate the group prior $P(y, a)$. The associated attribute value in Eq. 5 can be estimated as $a_{\mathbf{x}} = \arg\max_a P(a|\mathbf{x})$.

4.2.2 Many-to-One

Under this scenario, multiple labels can be correlating with the same attribute value. Without loss of generality, we assume that $y^{(1)}$ and $y^{(2)}$ are correlated with $a^{(1)}$, and $y^{(j)}, j > 2$ is correlated with $a^{(j-1)}$. We consider that the first 2 label predictions are spuriously correlated with attribute value $a^{(1)}$, *i.e.*, $\hat{f}_j(\mathbf{x}) \propto P(y^{(j)}, a^{(1)}|\mathbf{x}), j = 1, 2$ and other predictions are similar as the one-to-one situation, where $\hat{f}_j(\mathbf{x}) \propto P(a^{(j-1)}|\mathbf{x}), j > 2$. Compared to the one-on-one mapping, the only difference is the calculation of $P(a = a^{(1)}|\mathbf{x}_i)$ in Eq. 11. Considering both $y^{(1)}$ and $y^{(2)}$ contribute to $a^{(1)}$, we have,

$$\begin{aligned} P(a = a^{(1)}|\mathbf{x}_i) &= P(y^{(1)}, a = a^{(1)}|\mathbf{x}_i) + P(y^{(2)}, a = a^{(1)}|\mathbf{x}_i) \\ &= (\exp(\hat{f}_1(\mathbf{x}_i)) + \exp(\hat{f}_2(\mathbf{x}_i)))/\sum_{k=1}^{K} \exp(\hat{f}_k(\mathbf{x}_i)) \end{aligned} \tag{12}$$

The associated attribute value can then be estimated as well, *i.e.* $a_{\mathbf{x}} = \arg\max_a P(a|\mathbf{x})$.

4.2.3 One-to-Many

In this scenario, one label can be correlating with the multiple attribute values. Since we don't have the attribute label, to distinguish different attributes correlated with the same label, we follow Seo et al. (2022) to create pseudo labels for multiple attribute values. Without loss of generality, we assume $y^{(1)}$ is correlated with $a^{(1)}$ and $a^{(2)}$. Therefore, $P(a = a^{(j)}|\mathbf{x}_i) = \frac{w_j}{w_1+w_2} \exp(\hat{f}_1(\mathbf{x}_i))/\sum_{k=1}^{K} \exp(\hat{f}_k(\mathbf{x}_i)), j = 1, 2$, where w is the weight defined in Seo et al. (2022), Eq. (6). The associate attribute value can also be estimated with the estimated posterior.

4.2.4 Many-to-Many

Since this is the combination of the previous cases, we can apply solutions mentioned above together. In order to accurately calculate the prior probability $\mathbf{P}(y, a)$, we need to at least know how the label set $\mathcal{Y}$ and the attribute set $\mathcal{A}$ are correlated. Annotating the category-level relation is much easier than annotating the sample-level attribute. In the case where even the category-level relation is not available, we show in Sec. E that directly applying the one-to-one assumption in other situations (one-to-many and many-to-one) still shows reasonable performance.

4.3 Group MixUp

To further mitigate the geometric skew and increase the diversity of the samples in the minority group, we proposed a simple group MixUp method. We also start with the one-to-one situation and other situations can be derived similarly. Same as Sec. 4.2.1, without loss of generality, we assume the j-th category ($y^{(j)}$) is correlated with the j-th attribute value ($a^{(j)}$). An training sample $(\mathbf{x}_i, y_i)$ is in the minority group when $\arg\max_{y'} \hat{f}'_y(\mathbf{x}_i) \neq y_i$, else it is in the majority group. Given one sample $\mathbf{x}_i$ in the minority group, we randomly select one sample $\mathbf{x}_j$ in the majority group with the same label ($y_i = y_j$), instead of using the original $\mathbf{x}_i$ in training, following the idea of MixUp (Zhang et al., 2018), we propose to generate a new training example $(\mathbf{x}'_i, y_i)$ as well as its correction term via the linear combination of the two examples,

$$\mathbf{x}'_i = \lambda \mathbf{x}_i + (1 - \lambda)\mathbf{x}_j, \quad \Delta'_{y_i} = \lambda \Delta_{y_i} + (1 - \lambda)\Delta_{y_j}, \tag{13}$$

where $\lambda \sim U(0.5, 1)$ to assure that the mixed example is closer to the minority group example. Since both samples are with the same label, we expect their convex combination shares the same label. Using such a convex combination technique increases the diversity of minority groups.

5 EXPERIMENT

In this section, we evaluate the effectiveness of the proposed logit correction (LC) method on five computer vision benchmarks presenting spurious correlations: Colored MNIST (C-MNIST) (Arjovsky et al., 2020), Corrupted CIFAR-10 (C-CIFAR10) (Hendrycks & Dietterich, 2019; Nam et al., 2020), Biased FFHQ (bFFHQ) (Karras et al., 2019; Lee et al., 2021), Waterbird (Wah et al., 2011), and CelebA (Liu et al., 2015). Sample images in the datasets can be found in Figure 5 of Appendix.

5.1 EXPERIMENTAL SETUP

Datasets. C-MNIST, C-CIFAR-10 and Waterbird are synthetic datasets, while CelebA and bFFHQ are real-world datasets. The above datasets are utilized to evaluate the generalization of baselines over various domains. The C-MNIST dataset is an extension of MNIST with colored digits, where each digit is highly correlated to a certain color which constitutes its majority groups. In C-CIFAR-10, each category of images is corrupted with a certain texture noise, as proposed in Hendrycks & Dietterich (2019). Waterbird contains images of birds as "waterbird" or "landbird", and the label is spuriously correlated with the image background, which is either "land" or "water". CelebA and bFFHQ are both human face images datasets. On CelebA, the label is blond hair or not and the gender is the spurious attribute. The group containing samples of male with blond hair is the minority group. bFFHQ uses age and gender as the label and spurious attribute, respectively. Most of the females are "young" and males are "old".

Evaluation. Following Nam et al. (2020), for C-MNIST and C-CIFAR-10 datasets, we train the model with different ratios of the number of minority examples to the number of majority examples and then test the accuracy on a group-balanced test set (which is equivalent to GBA). The ratios are set to 0.5%, 1%, 2%, and 5% for both C-MNIST and C-CIFAR-10. For bFFHQ dataset, the model is trained with 0.5% minority ratio and the accuracy is evaluated on the minority group, following Lee et al. (2021). For Waterbird and CelebA datasets, we measure the worst group accuracy (Sohoni et al., 2020).

Baselines. We consider six baselines methods:

- *Emperical Risk Minimization (ERM)*: training with standard softmax loss on the original dataset.
- *Group-DRO* (Sagawa et al., 2019): Using the ground truth group label to directly maximize the worst-group accuracy.
- *GEOGRE* (Sohoni et al., 2020): Using clustering to generate pseudo group labels and then reweighting the minority group.
- *Learn from failure (LfF)* (Nam et al., 2020): Using ERM to detect minority samples and estimate a weight to reweight minority samples.
- *Just train twice (JTT)* (Liu et al., 2021b): Similar to LfF but weight the minority samples by a hyperparameter.
- *Disentangled feature augmentation (DFA)* (Lee et al., 2021): Using the generalized cross entropy loss (Zhang & Sabuncu, 2018) to detect minority samples and reweight the minority. Using feature swapping to augment the minority group.

Implementation details. We deploy a multi-layer perception (MLP) with three hidden layers as the backbone for C-MNIST, and ResNet-18 for the remaining datasets except ResNet-50 for Waterbirds and CelebA. The optimizer is Adam with $\beta = (0.9, 0.999)$. The batch size is set to 256. The learning rate is set to 1×10^{-2} for C-MNIST, 1×10^{-3} for Waterbird and C-CIFAR-10, and 1×10^{-4} for CelebA and bFFHQ. For q in Eq. 3, it's set to 0.7 for all the datasets except for Waterbird which is set to 0.8. More details are described in Appendix. D.

Table 1: Classification accuracy (%) evaluated on group balanced test sets of C-MNIST and C-CIFAR-10 with varying ratio (%) of minority samples. The baseline method results are taken from Lee et al. (2021) as the same experiment settings are adopted. We denote whether the model requires group or spurious attribute annotations in advance by ✗(i.e., not required), and ✓(i.e., required). Best performing results are marked in **bold**.

Methods	Group Info	C-MNIST				C-CIFAR-10			
		0.5	1.0	2.0	5.0	0.5	1.0	2.0	5.0
Group DRO	✓	63.12	68.78	76.30	84.20	33.44	38.30	45.81	57.32
ERM	✗	35.19	52.09	65.86	82.17	23.08	25.82	30.06	39.42
GEOGRE	✗	36.04	56.44	65.10	81.64	15.12	19.15	28.83	43.29
LfF	✗	52.5	61.89	71.03	80.57	28.57	33.07	39.91	41.78
DFA	✗	65.22	81.73	84.79	89.66	29.95	36.49	41.78	51.13
LC(ours)	✗	**71.25**	**82.25**	**86.21**	**91.15**	**34.56**	**37.34**	**47.81**	**54.55**

Table 2: Worst-group accuracies on Waterbirds, CelebA, and minority-group accuracy on bFFHQ. For the ERM, JTT and Group-DRO baselines, we provide the results reported in Liu et al. (2021a), except for bFFHQ, we rerun the baseline methods. The Group Info column shows whether group labels are available during training.

Method	Group Info	Waterbirds	CelebA	bFFHQ
		Worst	Worst	Minority
Group-DRO	✓	91.4	88.9	-
ERM	✗	72.6	47.2	55.47
LfF	✗	78.0	77.2	62.24
JTT	✗	86.7	81.1	65.31
LC(ours)	✗	**90.5**	**88.1**	**69.97**

6 Results

6.1 Classification Accuracy

Table 1 reports the accuracies on group balanced test sets for all baseline approaches and the proposed method when trained with different minority to majority ratios. Models trained with ERM commonly shows degraded performance and the phenomenon is aggravated as the decrease of the amount of examples in the minority groups. Compared to other approaches, LC consistently achieves the highest test accuracy. The performance gain is even more significant when the minority ratio is low. For example, compared to DFA Sohoni et al. (2020), LC improves the accuracy by 6.03%, 4.61% on C-MNIST and C-CIFAR-10 datasets respectively, when the minority ratio is 0.5%. While at 5% minority ratio, the improvements are 1.51% and 3.42%. It shows the superior performance of the proposed method in datasets with strong spurious correlation. Even compared to the approach which requires the ground-truth attribute label during training (Group DRO), LC still achieves a competitive or even better performance. Table 2 shows the model performances on Waterbird, CelebA and bFFHQ datasets. LC again achieves the highest worst/minority-group accuracy among all methods without group information. The clear performance gaps again prove the effectiveness of the proposed method.

Note that all results demonstrated in both Table 1 and 2 are on datasets with one-to-one spurious correlation. We also tested the proposed algorithm on datasets with many-to-one and one-to-many correlations as well. The proposed method also outperforms the best baseline methods. Additional experiments on can be found in Appendix E.

6.2 Ablation Study

Effectiveness of each module. Table 3 demonstrates the effectiveness of the LC loss and the Group MixUp in the proposed method. The evaluation is conducted on the bFFHQ dataset. The first row shows the performance of the baseline ERM network. From row 2-4, each proposed module helps improve the baseline method. Specifically, adding Group MixUp brings 6.35% of performance boost, and introducing logit correction is able to improve the performance by 9.64%. Combining both of the elements achieves 12.80% accuracy improvement. Compared our method without Group MixUp (third row in Table 3) to LfF in Table 2, both methods use the same pipeline and the only

difference is that we apply LC loss while LfF uses reweighting. Experiment result shows that the proposed LC loss clearly outperforms reweighting (62.2% $\rightarrow$ 66.51%). The reasons may due to the proposed LC loss 1) is Fisher consistent with the balanced-group accuracy; and 2) is able to reduce the geometric skew as well as the statistical skew.

Table 3: Ablation studies on 1) Group MixUp, 2) correcting logit on bFFHQ. Each row indicates a different training setting with ✓mark denoting the setting applied. While correcting the logit individually brings significant performance boost, adding Group MixUp further improves the performance.

Group MixUp	Logit Correction	Minority Group Accuracy
✗	✗	56.87
✓	✗	63.22
✗	✓	66.51
✓	✓	69.67

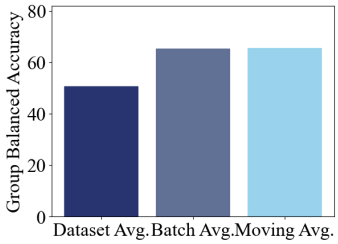
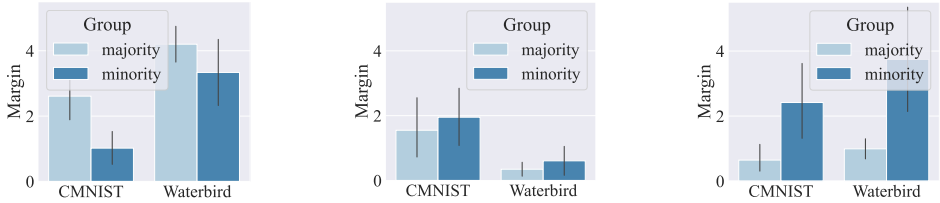


Figure 3: The performance comparison of different strategies for estimating the group prior (on CMNIST with ratio = 0.5%).

Influence of group prior estimate method. We test how different group prior estimation strategies in Eq. 9 affect the final performance. We tested 1) updating the prior using all samples in the datasets after finishing each training epoch (Dataset Avg.); 2) updating the prior using all samples in one training mini-batch (Batch Avg.); and 3) keeping a moving average for the batch-level prior (Moving Avg.). The result is shown in Fig. 3. Dataset Avg. performs significantly worse than other two strategies. This maybe because that the Dataset Avg. only updates the prior after each epoch. The delay in the prior estimation may mislead the model training especially in the early training stage when the model prediction can change significantly.



(a) ERM (b) ERM + Group MixUp (c) LC Loss

Figure 4: The effect of the proposed logit correction (LC) method on classification margins (defined in Appendix C) on CMNIST and Waterbird datasets. ERM produces a ratio (between the majority group margin and the minority group margin) > 1, ERM + Group Mixup has a ratio < 1 and the proposed LC loss achieves a ratio $\ll 1$.

Analysis of training margins. We show how the proposed LC loss and Group MixUp help reduce the geometric skew. In Sec. 1, we mentioned that a balanced classifier prefers a larger margin on the minority group comparing to the margin on the majority group, *i.e.*, the ratio between the majority group margin and the minority group margin should less than 1. In Figure 4, we show the minority group margin and the majority group margin (defined in Appendix C) of the model trained with ERM, LC loss and ERM + Group MixUp on both C-MNIST and Waterbird datasets respectively. Figure 4 shows that both the proposed LC loss and the Group MixUp can reduce the geometric skew since both of them have a ratio less than 1.

7 Conclusion

In this work, we present a novel method consisting of a logit correction loss with Group MixUp. The proposed method can improve the group balanced accuracy and worst group accuracy in the presence of spurious correlations without requiring expensive group labels during training. LC is statistically motivated and easy-to-use. It improves the group balanced accuracy by encouraging large margin for minority group and reducing both statistical and geometric skews. Through extensive experiments, The proposed method achieves the state-of-the-art group-balanced accuracy and worst-group accuracy across several benchmarks.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *stat*, 1050:27, 2020.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32, 2019.

Guillem Collell, Drazen Prelec, and Kaustubh Patil. Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data. *arXiv preprint arXiv:1606.08698*, 2016.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34:12091–12103, 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019.

Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14992–15001, 2021.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25123–25133, 2021.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9572–9581, 2019.

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

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 *Proceedings of the 38th International Conference on Machine Learning*, pp. 6781–6792, 2021b.

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

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pp. 603–611. PMLR, 2013.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*. OpenReview.net, 2021.

Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, pp. 6927–6937. PMLR, 2020.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Thao Nguyen, Vaishnavh Nagarajan, Hanie Sedghi, and Behnam Neyshabur. Avoiding spurious correlations: Bridging theory and practice. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 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.

Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10): 896–904, 2021.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

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. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16742–16751, 2022.

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

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

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 *International Conference on Machine Learning, ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26484–26516. PMLR, 2022.

Zhi Zhang, Tong He, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of freebies for training object detection neural networks. *arXiv preprint arXiv:1902.04103*, 2019.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.