APPENDIX

A PROOF OF PROPOSITION 1

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

$$y = \arg\max_{y'} \sum_{a} \frac{P(y',a|\mathbf{x})}{P(y',a)} = \arg\max_{y} \sum_{a} \frac{P(y|a,\mathbf{x})P(a|\mathbf{x})}{P(y,a)}$$

is the solution to Eq. 2

Proof. To simplify the notation, we define the output of the classifier as $c = \arg\max_{y' \in \mathcal{Y}} f_{y'}(\mathbf{x})$.

Following Collell et al. (2016), for a group $(y^{(j)}, a^{(k)})$, the accuracy in this group can be written as,

$$Acc(y^{(j)}, a^{(k)}) = \int_{\mathbf{x}} \frac{P(y=y^{(j)}, a=a^{(k)}|\mathbf{x})P(c=y^{(j)}|\mathbf{x})}{P(y=y^{(j)}, a=a^{(k)})} P(\mathbf{x}) d\mathbf{x}. \tag{14}$$

GBA in Eq. 2 thus can be rewritten as:

$$GBA = \frac{1}{KL} \int_{\mathbf{x}} \sum_{y^{(j)}} \sum_{a^{(k)}} \frac{P(y=y^{(j)}, a=a^{(k)}|\mathbf{x})P(c=y^{(j)}|\mathbf{x})}{P(y=y^{(j)}, a=a^{(k)})} P(\mathbf{x}) d\mathbf{x}. \tag{15}$$

Maximizing Eq. 15 is equivalent to obtain the optimal choice of $P(c=y^{(j)}|\mathbf{x})$ at each $\mathbf{x}$. Since what inside of the integral is

$$\begin{aligned} &\sum_{y^{(j)}} \sum_{a^{(k)}} \frac{P(y=y^{(j)}, a=a^{(k)}|\mathbf{x})P(c=y^{(j)}|\mathbf{x})}{P(y=y^{(j)}, a=a^{(k)})} \\ =&\sum_{y^{(j)}} \left(\sum_{a^{(k)}} \frac{P(y=y^{(j)}, a=a^{(k)}|\mathbf{x})}{P(y=y^{(j)}, a=a^{(k)})} \right) P(c=y^{(j)}|\mathbf{x}), \end{aligned} \tag{16}$$

which is a convex combination, and is maximized at each $\mathbf{x}$ if and only if we place probability 1 to the largest term. That is to say, at each $\mathbf{x}$, we assign 1 to $P(c=y^{(j)}|\mathbf{x})$ where $\sum_{a^{(k)}} \frac{P(y=y^{(j)}, a=a^{(k)}|\mathbf{x})}{P(y=y^{(j)}, a=a^{(k)})}$ is the largest term among all possible y values in $\mathcal{Y}$ and assigning 0 to other terms. Formally,

$$P(c=y^{(j)}|\mathbf{x}) = \begin{cases} 1, & \text{if } y^{(j)} = \arg\max_{y'} \sum_{a^{(k)}} \frac{P(y=y', a=a^{(k)}|\mathbf{x})}{P(y=y', a=a^{(k)})} \\ 0, & \text{Otherwise.} \end{cases} \tag{17}$$

The second equation can be derived by Bayes' theorem. □

B PSEUDO CODE OF THE PROPOSED ALGORITHM

We provide the pseudo code of the proposed logit correction and Group MixUp in Algorithm 1.

C DEFINITION OF CLASSIFICATION MARGIN

Let $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^k$ be a model that outputs k logits, following previous works (Koltchinskii & Panchenko, 2002; Cao et al., 2019), we define the margin of an example (x,y) as

$$m(x,y) = f(\mathbf{x})_y - \max_{j \neq y} f(\mathbf{x})_j. \tag{18}$$

We can then define the training margin for a group $g=(a,y)$ as the minimum margin of all classes

$$m_g = \min_{i \in g} m(x_i, y_i). \tag{19}$$

The margin for minority group and majority groups are defined as the average margin of minority/majority groups.

Algorithm 1: LC for one-to-one mapping

Input : Training set (X, Y), Initialize the ERM model $\hat{f}_\theta$ and the robust model f_θ, # epochs K, # rampup epoch T, moving average momentum α.

for *epoch* = 1 **to** K **do**
  Sample a mini-batch $\{(x, y)\}$;
  Update ERM network $\hat{f}(\theta)$ parameters by training on $\{(x, y)\}$ with Equation 3;
  for $(x, y) \in \{(x, y)\}$ **do**
    Let $p^{(x,y)}$ be the ERM model's probability outputs on sample (x, y).
    Let $\hat{a} := \arg\max p^{(x,y)}$
    Update the group priors by $\Delta_{y,\hat{a}} := \alpha\Delta_{y,\hat{a}} + (1-\alpha)p_{\hat{a}}^{(x,y)}$.
  end
  (Optional) Perform Group MixUp to obtain the synthesized batch:
  $\tau := 0.5 \cdot \exp(-5(1 - \textit{epoch})/T)^2$ sigmoid ramp up function
  $\{x, y, \Delta^{(x)}\} = \text{GroupMixup}(\{x, y, \hat{a}\}, \Delta, \tau)$
  for $(x, y, \Delta^{(x)}) \in \{x, y, \Delta^{(x)}\}$ **do**
    for l = 1 **to** L **do**
      Correct the lth logit of the robust mode by $f(x)_l := f(x)_l + \log \Delta_{l,\hat{a}}^{(x)}$
    end
  end
  Update robust model's parameters with softmax cross entropy loss on $f(x)$.
end
Function `GroupMixup` $(\{x, y, \hat{a}\}, \Delta, \tau)$**:**
  Obtain a set of samples$\{(\bar{x}, \bar{y}, \bar{a})\}$ that are estimated to be from minority groups i.e. $y \neq \hat{a}$
  $\{(\bar{x}, \bar{y}, \bar{a})\} := \text{shuffle}(\{(\bar{x}, \bar{y}, \bar{a})\})$;
  Sample $\lambda \sim \text{Uniform}(1 - 2\tau, 1 - \tau)$;
  for $i, (x, y, \hat{a}) \in$ *enumerate*$(\{x, y, \hat{a}\})$ **do**
    Sample $(\bar{x}, \bar{y}, \bar{a})$ from $\{(\bar{x}, \bar{y}, \bar{a})\}$ such that $y = \bar{y}$
    $x := \lambda x + (1 - \lambda)\bar{x}$
    $y := y$
    $\Delta^{(x)} := \lambda\Delta_y + (1 - \lambda)\Delta_y$ correction term for sample x.
  end
End Function

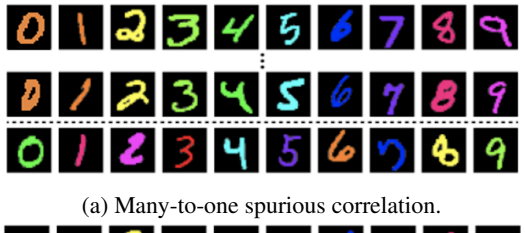
D Experiment Details

We utilize Adam optimizer with $\beta = (0.9, 0.999)$ without weight decay except for CelebA, we set weight decay to 1×10^{-4}, and a batch size of 256. For Waterbird, we use SGD optimizer with weight decay of 1×10^{-4}. Learning rates of 1×10^{-2}, 1×10^{-3} and 1×10^{-4} are used for Colored MNIST, Waterbird, and CelebA, respectively. We use a learning rate of 5×10^{-4} for 0.5% ratio of Corrupted CIFAR-10 and 1×10^{-3} for the remaining ratios. We decay learning rate at 10k iteration by 0.5 for both Colored MNIST and Corrupted CIFAR10. For CelebA, we adopt a cosine annealing learning rate schedule. For Waterbird, we set the q in GCE as 0.8 and 0.7 for other datasets. The rampup epoch T is set as 50 for waterbird and CelebA and 2 for other datasets, moving average momentum α is set to 0.5 for all datasets.

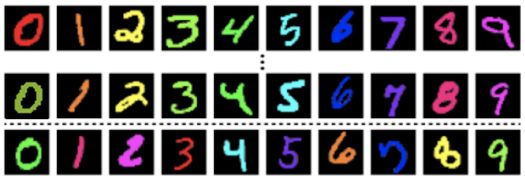
E Results on other Spurious Correlation

In the previous section, we report the results on one-to-one mapping which is a common scenario considered by previous works. In this section, we further examine the performance of previous approaches as well as LC on other other spurious correlation types.

Figure 5: Example images of datasets used in our work. In each dataset, the images above the dotted line demonstrate the majority groups while the ones below the dotted line are minority groups. For Colored MNIST and Corrupted CIFAR-10, each column demonstrates each class.



(a) Many-to-one spurious correlation.



(b) One-to-many spurious correlation.

Figure 6: Sample images for datasets containing one-to-many and many-to-one correlations. The first two rows show the majority groups' training samples and the third row shows the validation set where groups are balanced. For many-to-one, both digit 0's and 1's are colored by brown. For one-to-many, digit 0's are colored by different colors (red and dark green).

E.1 MANY-TO-ONE

The digits in MNIST training dataset is injected by different colors (similar to the original Colored MNIST). However, we inject the same color into digits 0 and 1 to obtain the many-to-one relationships between the label and the spurious attribute (Figure 6a). We evaluate the accuracy of the proposed logit correction method on many-to-one setting mentioned in Sec. 4.2.2 (LC+). We also apply LC loss with the one-to-one assumption as LC.

Table 4: Test accuracy on Colored MNIST data with many-to-one correlation

Methods	Group Info		Colored MNIST			
	Train	Val	0.5	1.0	2.0	5.0
ERM	✗	✓	31.13	50.89	57.92	82.19
LfF	✗	✓	46.22	69.32	73.33	82.57
DFA	✗	✓	64.7	77.28	84.19	90.17
LC(ours)	✗	✓	**65.32**	78.05	84.24	**90.3**
LC+(ours)	✗	✓	65.06	**78.57**	**84.5**	**90.3**

In Table 4, we report the accuracies on the balanced test set for Colored MNIST. The proposed method (LC and LC+) constaintly outperforms all baselines. Although using the exact mapping information shows the best performance (LC+), directly applying the one-to-one assumption shows very similar performance.

E.2 ONE-TO-MANY

We augmented the MNIST training dataset with colors (similar to Colored MNIST) except the Digits 0 has two colors as its major color attribute, as shown in 2. We evaluate the accuracy of the proposed logit correction method on one-to-many setting mentioned in Sec.4.2.3 (LC+). We also apply LC loss with one-to-one assumption as LC.

Table 5: **Benchmark results on One to Many correlation** Test accuracy on Colored MNIST data with one to many mapping

Methods	Group Info		Colored MNIST			
	Train	Val	0.5	1.0	2.0	5.0
ERM	✗	✓	38.47	48.41	67.41	80.61
LfF	✗	✓	52.79	66.07	75.09	83.5
DFA	✗	✓	70.11	78.75	83.06	90.41
LC(ours)	✗	✓	72.02	79.5	83.24	90.83
LC+(ours)	✗	✓	**72.26**	**80.1**	**84.1**	**91.25**

In Table 5, We report the accuracies on the balanced test set for Colored MNIST. The proposed method (LC and LC+) constaintly outperforms all baselines. Although using the exact mapping information shows the best performance (LC+), directly applying the one-to-one assumption shows very similar performance.