Controllable Prompt Tuning For Balancing Group Distributional Robustness

Hoang Phan [1] **Andrew Gordon Wilson** [1] **Qi Lei** [1]

Abstract

Models trained on data composed of different groups or domains can suffer from severe performance degradation under distribution shifts. While recent methods have largely focused on optimizing the worst-group objective, this often comes at the expense of good performance on other groups. To address this problem, we introduce an optimization scheme to achieve good performance across groups and find a good solution for all without severely sacrificing performance on any of them. However, directly applying such optimization involves updating the parameters of the entire network, making it both computationally expensive and challenging. Thus, we introduce Controllable Prompt Tuning (CPT), which couples our approach with prompt-tuning techniques. On spurious correlation benchmarks, our procedures achieve state-of-the-art results across both transformer and non-transformer architectures, as well as unimodal and multimodal data, while requiring only 0.4% tunable parameters.

1. Introduction

Spurious correlation or shortcut learning arises when a classifier relies on non-predictive features that coincidentally correlate with class labels among training samples. For example, a *majority* of waterbirds are often found near water, while landbirds mostly appear near land. Even when being provided with a small number of *minority group examples*, such as waterbirds on land, or landbirds on water, model predictions are often still dominated by spurious features, in this case, the background, ignoring the salient feature, the foreground (Sagawa et al., 2019).

A variety of approaches aim to improve worst group performance, by oversampling high-loss samples (Liu et al., 2021b; Zhang et al., 2022), undersampling the majority group (Sagawa et al., 2020) or re-training a last layer on group-balanced validation data (Kirichenko et al., 2022). A particularly popular and effective baseline, *Group Distributionally Robust Optimization* (GroupDRO) (Sagawa et al., 2019; 2020), directly minimizes an estimated upper bound on worst group loss. However, all these procedures generally neglect knowledge transfer between groups. Moreover, high-error samples can potentially include noisy-label data that could hurt the model's predictive ability if we try to learn them exhaustively (Oh et al., 2022). Besides, DRO-based methods in particular are susceptible to overfitting, wherein their test performances can decline to the same as ERM with sufficiently long training (Gulrajani & Lopez-Paz, 2020; Piratla et al., 2021; Zhai et al., 2022).

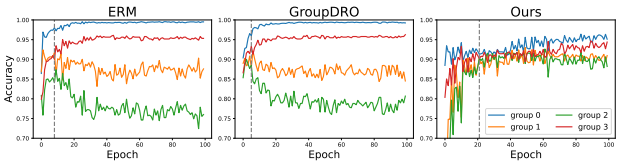


Figure 1: Accuracy curves on Waterbirds for four groups during training. Vertical lines indicate early stopping epochs as models obtain the best performance on the validation set.

To visually illustrate these limitations, we plot in Figure 1 the performance of a ResNet50 on the Waterbirds benchmark, trained with ERM, GroupDRO, and our approach (to be explained later). The plots demonstrate that while GroupDRO can improve the performance of the minority group (green line) early in training compared to ERM, it rapidly overfits and fails to maintain this performance over time. After training for ten epochs, the test performance gap between the minority and majority (blue line) groups grows sharply. To address these critical limitations, we introduce a systematic approach to train a model that exhibits consistently high performance across all groups. Specifically, at each iteration, we identify a descending direction that benefits the groups simultaneously rather than focusing on an individual group. Figure 1 shows that our proposed method performs almost equally well in every group, thanks to our balancing mechanism.

Moreover, we aim to control the group-wise learning process by ensuring that the magnitude of loss for different groups is inversely proportional to a predefined vector $\mathbf{c}$.

[1]New York University. Correspondence to: Hoang Phan <hvp2011@nyu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Varying this hyperparameter $\mathbf{c}$ adjusts the trade-off between the worst group and the average performance, yielding a more flexible method than prior work that only focuses on maximizing the worst group accuracy. However, computation can grow linearly with the number of controlling vectors $\mathbf{c}$. To overcome this challenge, we propose decoupling our optimization framework with parameter-efficient fine-tuning (PEFT) techniques that considerably reduce the number of trainable parameters. With the model compression ability powered by prompt-tuning, our proposed model can scale well for a large number of controlling vectors with a small parameter increase, e.g., 0.01% and 0.003% per value of $\mathbf{c}$ for Vision Transformer (Dosovitskiy et al., 2020) and CLIP (Radford et al., 2021), respectively.

Contribution. In this work, we introduce **CPT**, a **C**ontrollable **P**rompt **T**uning method to prevent models from learning spurious features. In summary, our key contributions are depicted as follows:

• Drawing inspiration from the principles of multi-objective optimization theory, we introduce a novel balancing mechanism to address the challenges of group distributional robustness. Our method not only considers the worst group but also leverages the gradient information from all groups to determine an ultimate updating direction that benefits all of them.

• While the above rigid balancing procedure works well in different scenarios, we extend our method by introducing a controlling vector that allows us to dynamically adjust the priority across groups. Additionally, we integrate prompt tuning techniques to enhance the scalability of our method as the complexity and number of controlling vectors increase and to make it applicable for de-biasing large transformers-based models.

• We conduct intensive experiments on many different benchmarks, where CPT consistently exhibits superior performance. Our method surpasses the current state-of-the-art baselines on Waterbirds andimprove the performance CelebA datasets while updating 0.4% parameters. Moreover, it also outperforms recent proposed methods that aims to de-bias Vision Transformer and CLIP models with minimal training cost.

2. Related work

We first review relevant methods to enhance distributional robustness and then discuss prior work related to the two essential parts of our proposed method: transformer and parameter-efficient fine-tuning.

Distributional robustness can be compromised for reasons like selection bias or spurious correlation across tasks. For selection bias on individual samples, prior works seek to reduce bias via practices like importance reweighting (Heckman, 1979; Shimodaira, 2000; Cortes et al., 2010; Lei et al., 2021), hard sample reweighting (Liu et al., 2021b; Nam et al., 2020) or distribution matching and discrepancy minimization (Cortes et al., 2015; Ben-David et al., 2010; Berthelot et al., 2021), and domain-adversarial algorithms (Ganin et al., 2016; Long et al., 2018; Phan et al., 2023) across tasks in their feature representation space (Tran et al., 2023). For selection bias on groups or subpopulation, namely subpopulation shift or dataset imbalance, label propagation (Cai et al., 2021; Berthelot et al., 2021) or other consistency regularization (Miyato et al., 2018; Yang et al., 2023a) are used to generalize the prediction to broader domains.

Improving distributional robustness also requires mitigating the effect of environmental features. Prior work uses model ensemble (Kumar et al., 2022) or model soups (Wortsman et al., 2022) to learn rich features and reduce the effect of spurious features. More practices include invariant feature representation learning (Arjovsky et al., 2019; Chen et al., 2022b), domain generalization (Krueger et al., 2021; Sun et al., 2017; Shi et al., 2021) or through data augmentation (Xu et al., 2020; Yao et al., 2022).

Transformer. In the past few years, transformer (Vaswani et al., 2017) has emerged as a dominant architecture for a wide range of applications, exhibiting their profound capacity in modeling different modalities: languages (Devlin et al., 2018; Liu et al., 2019; Touvron et al., 2023), vision (Dosovitskiy et al., 2020; Chen et al., 2021) and speech (Radford et al., 2023; Wang et al., 2023). Due to its ability to process multiple modalities, transformer-based architectures were widely adopted for image understanding using different types of supervision. For example, Vision Transformer (ViT) (Dosovitskiy et al., 2020) utilizes masked patches for pre-training and image-label supervision for fine-tuning. Notably, Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), which was pre-trained on billion-scale image-text pairs, has demonstrated excellent zero-shot performance in image classification tasks. Despite obtaining impressive results on various downstream tasks, over-parameterization is still a crucial problem for transformer (Fan et al., 2019; Panahi et al., 2021) which potentially causes overfitting (Li et al., 2023) and unintended biases (Agarwal et al., 2021; Wang et al., 2021; Du et al., 2022; Zhang & Ré, 2022) more severely than small-scale architectures (Sagawa et al., 2020).

Parameter-Efficient Fine-Tuning. While conventional training updates the entire network, prior work has shown that it is possible to achieve performance comparable with full fine-tuning by instead updating a small set of parameters (Lialin et al., 2023; He et al., 2023; 2021). Early work in this line of research includes adapter (Houlsby et al., 2019; Chen et al., 2022a; Sung et al., 2022), which adds

a lightweight fully connected network between the layers of a frozen pre-trained model. Since then, there have been many other methods that update subsets of the network (Guo et al., 2021; Gheini et al., 2021; Zaken et al., 2022) or reparametrize the weights of the network using low-rank decompositions (Aghajanyan et al., 2020; Hu et al., 2021; Edalati et al., 2022). Another appealing PEFT strategy is prompt tuning, which prepends a few trainable tokens before the input sequence or hidden state to encode task-specific knowledge to the pre-trained model (Lester et al., 2021; Li & Liang, 2021; Zhu et al., 2023b; Huang et al., 2023).

Gradient-based multi-task learning. Prior work often searches for an update direction that benefits all tasks (Yu et al., 2020; Liu et al., 2021a; Phan et al., 2022a) or controls the trade-offs between per-task performance (Lin et al., 2019; Mahapatra & Rajan, 2020; Phan et al., 2022b). Therefore, casting the problem of learning from multiple domains into a multi-task problem (Liang et al., 2021; Kim et al., 2023) allows us to apply results from multi-objective optimization theory. Building on these insights, we propose a method that prevents the dominance of any single domain.

3. Background

In this section, we first formulate the problem of group robustness and recap the technical details of GroupDRO, then revisit some concepts of multi-objective optimization.

3.1. Problem formulation

Formally, we consider the conventional setting of classifying an input $x \in \mathcal{X}$ as a target $y \in \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y}$ are input and label spaces, respectively. We are given a training dataset composed of K groups from the set $\mathcal{G}$, where each group $g \in \mathcal{G}$ consists of n_g instances sampled from the probability distribution $P_g(\mathcal{X}, \mathcal{Y})$. Since the numbers of examples are different among groups in the training data, we consider groups with relatively large n_g as majority groups and those with small n_g as minority groups. Our main goal is to develop a model that performs effectively across all groups within $\mathcal{G}$ to avoid learning spurious correlations.

Following previous methods (Sagawa et al., 2020), we adopt two metrics: worst group accuracy, indicating the minimum test accuracy across all groups $g \in \mathcal{G}$, and average accuracy, which represents the weighted average test accuracy with the weights corresponding to the relative proportions of each group in the training data. It is worth noting that commonly used datasets for group robustness often exhibit skewed training set distributions, the weighted average score is thus dominated by the performance of the model on groups that include large numbers of training data. Therefore, we propose to evaluate the model performance by additionally reporting the mean (unweighted average) accuracy score across different groups.

3.2. Group distributionally robust optimization

GroupDRO learns the model parameters θ by directly optimizing for the worst-group training loss as follows:

$$\max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim P_g(\mathcal{X},\mathcal{Y})} \left[\ell\left(f_\theta(x), y\right)\right]$$

for some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. Then, their updating formula at each step is given by:

$$g^* = \arg\max \ell_g(\theta) \quad \text{and} \quad \theta^{t+1} = \theta^t - \eta \nabla_\theta \ell_{g^*},$$

where η is the learning rate.

3.3. Multi-objective Optimization

Assume that we are given m objectives functions $f_i(x)$, $i \in [m]$ where $x \in \mathbb{R}^d$ and each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a scalar-valued function (we use the $[m]$ notation to denote the set $\{1, 2, \ldots, m\}$) . In the multi-objective optimization (MOO) problem, we are interested in minimizing a vector-valued objective function whose i-th component is $f_i(x)$:

$$\vec{F}(x) = [f_1(x), f_2(x), \ldots, f_m(x)] \tag{1}$$

In general, an optimal solution for this objective vector function does not exist since each of the individual function $f_i(x)$ does not necessarily guarantee to have the same minimum solution. Alternatively, we expect to obtain a solution, from which we cannot improve any specific objective without hurting another. According to the above formulation, the solution (Zitzler & Thiele, 1999) of the multi-objective minimization problem is formally defined as follows:

Definition 3.1. (Pareto dominance) Let x_1, x_2 be two solutions for the multi-objective optimization problem in Equation 1, x_1 dominates x_2 ($x_1 \prec x_2$) if and only if $f_i(x_1) \leq f_i(x_2) \forall i \in [m]$ and $\exists j \in [m]$ s.t. $f_j(x_1) < f_j(x_2)$.

Definition 3.2. (Pareto optimality) A solution x^* in problem 1 is said to be Pareto optimal if it is not dominated by any other solution. Therefore, the solutions set of the multi-objective minimization problem is given by $\mathcal{P} := \{x \mid \nexists x' : x' \prec x\}$. This set of all Pareto optimal solutions is called the Pareto set while the collection of their images in the objective space is called the Pareto front.

Gradient-based MOO. Désidéri (2012) shows that the descent direction can be found in the convex hull composed of gradients corresponding to different objectives $\{\boldsymbol{h}^i := \nabla f_i\}_{i=1}^m$:

$$\mathcal{CH} = \left\{ \boldsymbol{H}^\top \boldsymbol{\alpha} = \sum_{i=1}^m \alpha_i \boldsymbol{h}^i \mid \alpha \in \Delta_m \right\}$$

where Δ_m denotes the $m-1$ dimension simplex.

Moreover, they introduce the Multiple Gradient Descent Algorithm (MGDA), which calculates the minimum-norm

gradient vector $\boldsymbol{h}$ that lies in the convex hull: $\boldsymbol{h} = \text{argmin}_{h \in \mathcal{CH}} ||h||_2$. This approach can guarantee that the obtained solutions lie on the Pareto front, from which we cannot find an updating direction that decreases all objectives simultaneously.

4. Proposed method

In GroupDRO, the updating formula solely focuses on the worst-performing group, which is susceptible to challenges when groups exhibit varying levels of loss magnitude, noise, and transfer characteristics. For example, among the high-loss group, there might exist a few outliers that have large input shifts or even wrong labels, which can raise the training loss of the whole group and prevent the model from fitting other groups.

4.1. Balancing group distributional robustness

In this paper, we go beyond the worst group approach of GroupDRO to examine all groups at a time. To encourage the model to learn from different groups at each step to avoid struggling on just the hardest one only, we propose to minimize the following K-dimension loss vector:

$$\vec{\mathcal{L}}(\theta) = \left\{ \mathbb{E}_{(x,y)\sim P_{g^k}(\mathcal{X},\mathcal{Y})} \left[\ell\left(f_\theta(x), y\right) \right] \right\}_{k=1}^{K} \tag{2}$$

where K is the number of groups and g^k is the k-th group.

At t-th iteration, we search for an update vector that minimizes the weighted sum of group losses:

$$d_t(w) = w^\top \nabla \vec{\mathcal{L}}(\theta_t) = \sum_{k=1}^{K} w_k \nabla \ell_k(\theta_t) \tag{3}$$

where $\ell_k(\theta) = \mathbb{E}_{P_{g^k}} \left[\ell\left(f_\theta(x), y\right) \right]$ and $w \in \Delta_K$.

Then, the updating formula of our algorithm is then given by $\theta_{t+1} = \theta_t - \eta d_t(w)$ for some step size $\eta \in \mathbb{R}^+$. The coefficient w is chosen such that our update rule can benefit all groups $d_t(w)^\top \nabla \ell_k(\theta_t) \geq 0$. To this end, a straightforward approach that is readily applicable to optimize the objective (2) is applying MGDA to find the solution of (3) that yields the minimum Euclidean norm of the composite gradient:

$$d_t^* = \underset{w \in \Delta_K}{\arg\min} \left\| \sum_{k=1}^{K} w_k \nabla \ell_k(\theta_t) \right\|_2 \tag{4}$$

However, MGDA is often biased toward objectives with smaller gradient magnitudes (Liu et al., 2020), and lacks controllable property (i.e. could not adjust the group that we want to focus more or less at each update), which can be utilized to balance the loss between groups. Hence, we introduce a new objective that steers the updating direction to obtain our desired balanced magnitude among groups. Assuming that the entropy $H(p)$ measures the diversity of a continuous distribution p, this behavior could be achieved via maximization of the entropy with respect to the distribution of loss functions:

$$\begin{aligned} \mathcal{L}_{ent}(\theta) &= H\left(\text{softmax}\left(\vec{\mathcal{L}}(\theta)\right)\right) \\ &= H\left(\text{softmax}\left(\left\{ \mathbb{E}_{P_{g^k}} \left[\ell\left(f_\theta(x), y\right) \right] \right\}_{k=1}^{K}\right)\right) \end{aligned} \tag{5}$$

Theorem 4.1. *Assume that the loss function ℓ is differentiable up to the first order with respect to θ, then following*

$$d_{ent} := \sum_{i=1}^{K} \nabla \ell_i(\theta) \left[p_i \log(p_i) - p_i \sum_{j=1}^{K} \log(p_j) p_j \right]$$

where $p_i = \frac{e^{\ell_i(\theta)}}{\sum_{j=1}^{K} e^{\ell_j(\theta)}}$, maximizes the objective $\mathcal{L}_{ent}(\theta)$.

The proof of Theorem 4.1 is deterred to Appendix A. Theorem 4.1 provides us with the updating rule that increases the entropy term in Equation (5) the most. To balance the learning across groups, we seek the coefficient w that maximizes the cosine similarity with d_{ent} and does not conflict against the gradient vector of each group. Denotes $w_i' = p_i \log(p_i) - p_i \sum_{j=1}^{K} \log(p_j) p_j$, this is equivalent to optimizing:

$$\begin{aligned} w^* &= \underset{w \in \Delta_K}{\arg\max}\, d_t(w)^\top d_{ent} \\ &\text{s.t.} \quad d_t(w)^\top \nabla \ell_k(\theta_t) \geq 0 \;\forall\, k \in [K]. \end{aligned} \tag{6}$$

While the balance between loss magnitude is not achieved, we strictly follow the updating direction that decreases training errors of high-loss groups, but slightly relax constraints in the optimization problem (6) to allow bounded increases among groups that already have small loss magnitudes.

$$\begin{aligned} w^* &= \underset{w \in \Delta_K}{\arg\max}\, d_t(w)^\top d_{ent} \\ &= \underset{w \in \Delta_K}{\arg\max}\, w^\top \nabla \vec{\mathcal{L}}(\theta)^\top \nabla \vec{\mathcal{L}}(\theta) w' \end{aligned} \tag{7}$$

$$\begin{aligned} &\text{s.t.} \quad \forall\, k \in [K]: \\ &d_t(w)^\top \nabla \ell_k(\theta_t) \geq 0 \text{ if } \ell_k(\theta_t) = \max_{k \in [K]} \ell_k(\theta_t). \\ &d_t(w)^\top \nabla \ell_k(\theta_t) \geq d_{ent}^\top \nabla \ell_k(\theta_t) \text{ if } \ell_k(\theta_t) < \max_{k \in [K]} \ell_k(\theta_t). \end{aligned}$$

The second constraint guarantees that the obtained solution is not worse than d_{ent} on those groups in which the model is performing well or when d_{ent} itself intrinsically supports learning them. Once the loss magnitude is almost balanced ($||d_{ent}|| \leq \epsilon$), we simply update the model parameter by the average group gradients, i.e. $w = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$.

In summary, the key difference between our proposed method and GroupDRO is that we solve a small K-dimensional optimization problem to find the updating direction that not only focuses on the worst group but also

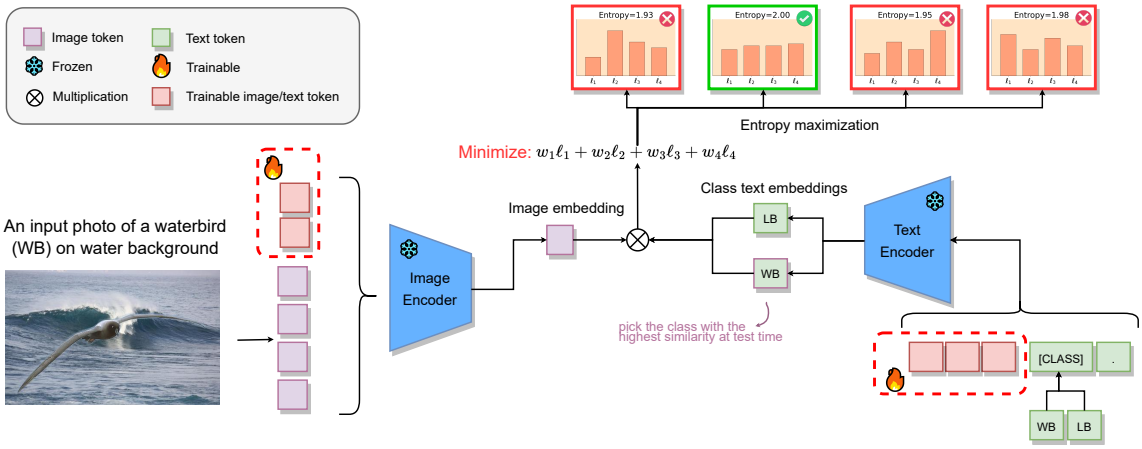


Figure 2: Overview of our method on the Waterbirds dataset. Our main objective is to not only improve model performance across groups by optimizing their loss functions $\ell_1, \ell_2, \ell_3, \ell_4$, but also maximize the entropy over this loss distribution.

improves the performance in other groups and balances the loss magnitude between them at the same time. In practical settings, the group number K is often a small number (e.g. $K = 4$), thus could be efficiently solved using standard linear programming libraries (Diamond & Boyd, 2016; Andersen et al., 2020) without introducing significant computational overhead.

4.2. Controllable prompt tuning for balancing group distributional robustness

To achieve the controllable property, we apply the optimization procedure in section 4.1 on $\ell_k(\theta) = c_k \mathbb{E}_{P_{g^k}} [\ell(f_\theta(x), y)]$ where $\mathbf{c} \in \mathbb{R}^K$ is a pre-defined vector that is used to adjust the magnitude of group losses. Intuitively, the principle behind our proposed objective is to enforce the loss vector $\vec{\mathcal{L}}(\theta)$ to be inversely proportional to $\mathbf{c}$, from which we can adjust the group loss magnitude, and thus control the trade-off between them by varying $\mathbf{c}$. For example, we can put more weight on penalizing loss terms of some specific groups to accelerate the learning progress on those groups, or conversely slow down the learning to avoid overfitting on oversampled groups (Deng et al., 2023).

In practice, tuning the value of $\mathbf{c}$ to obtain our desired model behavior requires learning multiple models individually and independently. Therefore, the computational complexity of our proposed method can grow linearly with the number of vectors $\mathbf{c}$, thus being very expensive in the context of deep learning. To reduce overhead, we adopt the usual practice of parameter-efficient fine-tuning methods and optimize only a small portion of the model. In particular, we freeze the main backbone while introducing lightweight prompt sets, one for each task. For example, we employ the same ViT backbone for the two datasets Waterbirds, CelebA, but using different sets of prompts (Section 5).

Overview of our proposed method is depicted in Figure 2, given an input image belonging to one of a group $\{g^1, g^2, \ldots, g^K\}$ and a set of classes, we solve the optimization problem (7) to find the reweighting coefficient w that maximizes the entropy $\mathcal{L}_{ent}(\theta)$ while reasonably decreases group losses. Please note that while Figure 2 illustrates our proposed method on the CLIP model on the image classification task, it is applicable to other transformer-based models and task types. In our experiment, we exploit both image-end prompts and language-end prompts tuning on frozen ViT and CLIP backbones, respectively.

Prompt tuning for Vision Transformer: Regarding the prompt design for a ViT of N layers, we use the set of continuous learnable tokens $\{\mathbf{P}_n\}_{n=0}^{N-1}$. Each $\mathbf{P}_n \in \mathbb{R}^{L \times D}$, where L, $D \in \mathbb{N}$ indicate the prompt length and latent space dimension, respectively. Subsequently, the input at the $(n+1)$-th layer follows the format of $[x_n, \mathbf{P}_n, \mathbf{E}_n]$, where $\mathbf{x}_n, \mathbf{E}_n \in \mathbb{R}^D$ represent the [CLASS] token and image patch embeddings after n-th layer while $[\cdot, \cdot]$ denotes the concatenation operator. Thus, the number of trainable parameters for this image-end prompt set is $N \times L \times D$.

Prompt tuning for CLIP: Being trained on an enormous amount of image-text pairs, CLIP offers an expressive representation for both language and image inputs by using two separate encoders, one for each modality, as shown in Figure 2. Since the image encoder can be either ResNet (He et al., 2016), ConvNeXt (Liu et al., 2022) or ViT (Dosovitskiy et al., 2020), we conduct prompt tuning on the transformer text encoder for consistency. In short, we employ a learnable context prompt of length L: $\mathbf{P} \in \mathbb{R}^{L \times D}$, to replace the hand-crafted prompt used by the original CLIP (i.e., "this is a photo of a "). Based on this design, the input for the text encoder of each class has the format of $[x, \mathbf{P}]$, where x is the [CLASS] embedding of the corresponding class name and has the same dimension of D. These learnable tokens $\mathbf{P}$ are better at capturing domains-specific knowledge than those artificial prompt (Zhou et al., 2022; Zhu et al., 2023a)

and can be learned to de-bias the original CLIP model while introducing a few trainable parameters ($L \times D$).

5. Experiments

In this section, we evaluate the effectiveness of the proposed CPT method on benchmark image datasets in the presence of spurious features: Waterbirds (Sagawa et al., 2019), CelebA (Liu et al., 2015), MetaShift (Liang & Zou, 2021) and ISIC (Codella et al., 2019). Due to the space limit constraint, we briefly provide an overview of our experimental setup below, see Appendix B for more details and additional results. Our implementation is available at `https://github.com/VietHoang1512/CPT`.

5.1. Datasets

Waterbirds is created by placing bird photos from the Caltech-UCSD Birds dataset (Wah et al., 2011) with background images taken from Places (Zhou et al., 2017). The target attributes are foreground objects (i.e. landbird or waterbird) while spurious correlations are background (i.e. land or water landscape). **CelebA** is a factual dataset containing 200K face images of celebrities. The goal is to classify blond/non-blond hair color. Statistically, more than 94% of blond hair examples in the training set are women.

MetaShift is another real-world dataset for cat-dog classification, where the cat and dog images are spuriously correlated with indoor (e.g. bed) and outdoor objects (e.g. bike). For evaluation, models are given dog and cat images associated with shelf backgrounds, which do not appear in the training set. Thus, this dataset represents a more coherent distributional shift, compared to other benchmarks. **ISIC** is an even more challenging dataset with the appearance of multiple spurious features (Codella et al., 2019), which includes dermoscopic images of skin lesions with target attributes benign or melanoma. Visual examples of those datasets, along with the size for each group in train, validation, and test sets are given in Appendix B.

5.2. Baselines

Following the previous studies, we compare CPT against other state-of-the-art methods for mitigating spurious correlations, including DFR (Kirichenko et al., 2022), Group-DRO (Sagawa et al., 2019), PDE (Deng et al., 2023), Subsample (Deng et al., 2023), CGD (Piratla et al., 2021) with the access to the group information during training. Besides, we consider those methods require group labels for validation: LfF (Nam et al., 2020), JTT (Liu et al., 2021b), PGI (Ahmed et al., 2020), CIM (Taghanaki et al., 2021), CNC (Zhang et al., 2022), or those do not use group annotations at all like ERM and AFR (Qiu et al., 2023). Moreover, we also report the results from two recent studies of spurious correlation on ViT (Ghosal & Li, 2023) and CLIP (Yang et al., 2023b; Dehdashtian & Boddeti, 2024).

Regarding the baselines for those experiments further taking the distribution shift into account, we follow exactly the protocol of Wu et al. (2023b) and compare our method to Upweighting (UW), IRM (Arjovsky et al., 2019), IB-IRM (Ahuja et al., 2021), V-REx (Krueger et al., 2021), CORAL (Sun et al., 2017), and Fish (Shi et al., 2021); instance reweighting methods: JTT (Liu et al., 2021b), DM-ADA (Xu et al., 2020), LISA (Yao et al., 2022) and the current state-of-the-art method, DISC (Wu et al., 2023b).

5.3. Model training and evaluation

We examine our proposed method on ResNet50 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2020) and different variants of OpenAI's CLIP (Radford et al., 2021) to align with previous work (Kirichenko et al., 2022; Ghosal & Li, 2023; Yang et al., 2023b). For prompt tuning on ViT, we employ prompts of lengths $L = 5$ in the MetaShift and ISIC, or $L = 10$ in the Waterbirds and CelebA datasets. Similarly, L is set to 16 for all CLIP backbones throughout our experiment. Unless otherwise explicitly mentioned in the experiment, the value of the controlling vector c is set to the default value of $(1, \dots, 1)$. Other detailed configurations are relegated to Appendix B.

Due to the large number of groups in ISIC, we calculate the AUROC score (Bradley, 1997), following (Wu et al., 2023b; Bissoto et al., 2020) while computing average, worst-group and mean (if possible) accuracy scores for other datasets. For a fair and precise comparison, results of baselines are taken from recent studies (Deng et al., 2023; Wu et al., 2023b; Piratla et al., 2021; Zhang et al., 2022; Kim et al., 2023) and reported directly from original papers.

5.4. Results

Waterbirds and CelebA. We start by validating CPT on two well-established benchmarks for investigating spurious correlation, namely Waterbirds and CelebA in Table 1. From the results, we see that CPT consistently achieves the best worst-group accuracy compared with the recent state-of-the-art methods such as DFR and PDE on both datasets. Remarkably, our method only updates a tiny portion (0.4%) of parameters. As such, CPT is much more computationally efficient than methods that need to update the entire network. Furthermore, we see that CPT has strong gains in average accuracy, thus not only outperforming other spurious correlation methods on both metrics but also closing the gap to the average accuracy of ERM. Another interesting observation is that those baselines aim to balance the learning between groups are effective in alleviating spurious correlation, e.g. via downsampling the majority groups (Subsample), reweighting group losses inversely proportional to

Table 1: Overall results on Waterbirds and CelebA datasets with best methods are highlighted in **bold**. Performance is evaluated on the test set with models early stopped at the highest worst-group accuracy on the validation set.

Method	Train group info	Validation group info	Train once	Waterbirds			CelebA		
				Worst	Average	# Params	Worst	Average	# Params
ERM	×	×	√	$70.0_{\pm 2.3}$	$97.1_{\pm 0.1}$	23M	$45.0_{\pm 1.5}$	$94.8_{\pm 0.2}$	23M
AFR	×	×	×	$90.4_{\pm 1.1}$	$94.2_{\pm 1.2}$	23M	$82.0_{\pm 0.5}$	$91.3_{\pm 0.3}$	23M
LfF	×	√	×	$78.0_{N/A}$	$91.2_{N/A}$	23M	$77.2_{N/A}$	$85.1_{N/A}$	23M
SSA	×	√	×	$89.0_{\pm 0.6}$	$92.2_{\pm 0.9}$	23M	$89.8_{\pm 1.3}$	$92.8_{\pm 1.3}$	23M
JTT	×	√	×	$86.7_{N/A}$	$93.3_{N/A}$	23M	$81.1_{N/A}$	$88.0_{N/A}$	23M
PGI	×	√	×	$79.5_{\pm 1.9}$	$95.5_{\pm 0.8}$	23M	$85.3_{\pm 0.3}$	$87.3_{\pm 0.1}$	23M
CIM	×	√	×	$77.2_{N/A}$	$95.6_{N/A}$	23M	$83.6_{N/A}$	$90.6_{N/A}$	23M
CnC	×	√	×	$88.5_{\pm 0.3}$	$90.9_{\pm 0.1}$	23M	$88.9_{\pm 1.3}$	$88.9_{\pm 0.5}$	23M
DFR	×	√	×	$92.9_{\pm 0.2}$	$94.2_{\pm 0.4}$	23M	$88.3_{\pm 1.1}$	$91.3_{\pm 0.3}$	23M
UW	√	√	√	$88.0_{\pm 1.3}$	$95.1_{\pm 0.3}$	23M	$83.3_{\pm 2.8}$	$92.9_{\pm 0.2}$	23M
Subsample	√	√	√	$86.9_{\pm 2.3}$	$89.2_{\pm 1.2}$	23M	$86.1_{\pm 1.9}$	$91.3_{\pm 0.2}$	23M
LISA	√	√	×	$89.2_{\pm 0.6}$	$91.8_{\pm 0.3}$	23M	$89.3_{\pm 1.1}$	$92.4_{\pm 0.4}$	23M
GroupDRO	√	√	√	$86.7_{\pm 0.6}$	$93.2_{\pm 0.5}$	23M	$86.3_{\pm 1.1}$	$92.9_{\pm 0.3}$	23M
CGD	√	√	√	$88.9_{\pm 0.8}$	$91.3_{\pm 0.6}$	23M	$90.0_{\pm 0.8}$	$92.5_{\pm 0.2}$	23M
DISC	√	√	×	$88.7_{\pm 0.4}$	$93.8_{\pm 0.7}$	23M	N/A	N/A	N/A
PDE	√	√	√	$90.3_{\pm 0.3}$	$92.4_{\pm 0.8}$	23M	$91.0_{\pm 0.4}$	$92.0_{\pm 0.6}$	23M
CPT	√	√	√	$\mathbf{93.5}_{\pm 0.4}$	$96.3_{\pm 0.2}$	**94**k	$\mathbf{92.0}_{\pm 0.3}$	$93.2_{\pm 0.3}$	**94**k

the group populations (UW), or computing scaling weights using group training performance to upweight the worst-group loss (GroupDRO). Even though, CPT still exhibits impressive performance compared to the aforementioned methods by using an adaptive balancing mechanism at each iteration instead of relying on fixed reweighting coefficients.

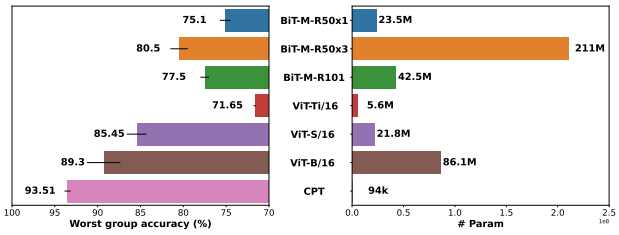


Figure 3: Results of fine-tuning ViT backbones on Waterbirds. Error bars represent the standard deviation over independent runs.

De-bias ViT and CLIP. Motivated by recent studies focusing on investigating spurious correlations of pre-trained transformer-based models (Yang et al., 2023b; Dehdashtian & Boddeti, 2024), we apply our proposed algorithm on ViT and CLIP and then compare them against those work. It is worth noting that, the authors did not experiment on CelebA with the blond hair target. Hence, we only show their original results on Waterbirds in Figure 3 and Table 2. As can be seen in Figure 3, while CPT enjoys much less computation and memory overhead (only 0.1% and 0.04% compared to ViT-B/16 itself and BiT-M-R50x3 (Ghosal & Li, 2023)) it still outperforms the second-best model by a large margin. This significant improvement comes from our prompting design and optimization contribution, respectively. Details of backbones are given in Appendix B.

Table 2: Performance of CLIP models on Waterbirds. † indicates our implemented results. We use **Bold** font to indicate the best methods for different numbers of trainable parameters. N/A denotes those methods do not report the number of trainable parameters.

Model	ResNet50			ViT-L/14@336px		
	Average	Worst	# Params	Average	Worst	# Params
Pre-trained CLIP	90.8	44.9	0	88.5	34.0	0
Fine-tuned CLIP	81.3	77.1	14789632	97.2	89.7	786432
ERM	93.5	54.4	14789632	96.8	58.1	786432
ERM Adapter	96.0	63.0	524288	97.8	76.1	524288
Contrastive Adapter	88.2	82.5	263424	94.5	85.3	197632
DFR	91.8	63.9	N/A	96.1	65.9	N/A
FairerCLIP	84.3	75.4	N/A	92.2	86.0	N/A
GroupDRO	83.3	73.7	14789632	94.1	90.8	786432
Yang et al. (2023b)	83.2	77.5	14789632	96.9	90.5	786432
$\text{CPT}_{\text{balance}}$	85.8	**82.8**	14789632	94.0	**92.9**	786432
ERM†	85.7	64.6	8192	97.4	84.3	12288
GroupDRO†	81.0	76.7	8192	95.8	90.5	12288
CPT	83.3	78.4	8192	97.0	90.2	12288
$\text{CPT}_{\text{balance}}$	81.8	**79.8**	8192	95.6	**91.9**	12288

As shown in Table 2, $\text{CPT}_{\text{balance}}$ (i.e. CPT with $c = (1, \ldots, 1)$) clearly outperforms all other alternatives using the same number of trainable parameters, which proves the effectiveness of our optimization framework. Furthermore, when varying the coefficient c, one can control the trade-off between group performance, yielding a much more flexible learning algorithm for group distributional robustness. Indeed, increasing coefficients of the majority groups (e.g. $c = (2, 1, 1, 1.5)$ in this setting) helps improve the average accuracy of CPT to perform on par or better than Yang et al. (2023b) on both metrics while updating only 0.05% and 1% of the model parameters. Thus, this again suggests the clear efficiency advantage of our prompt tuning, which offers scalability for controlling adjustment. Since CPT has a larger capacity to control the learning between groups, we can

sacrifice worst-group accuracy for the gain in performance on other groups and vice-versa (last two rows). Intriguingly, while obtaining impressive performance on CLIP-RN50, Contrastive Adapter (Zhang & Ré, 2022) falls far behind other baselines on CLIP ViT-L/14@336px.

Table 3: Experimental results on MetaShift and ISIC dataset.

	MetaShift		ISIC	
Method	Average	Worst	AUROC	# Params
ERM	$72.9_{\pm 1.4}$	$62.1_{\pm 4.8}$	$36.4_{\pm 0.7}$	23M
UW	$72.1_{\pm 0.9}$	$60.5_{\pm 3.8}$	$39.2_{\pm 0.6}$	23M
IRM	$73.9_{\pm 0.8}$	$64.7_{\pm 2.1}$	$45.5_{\pm 3.6}$	23M
IB-IRM	$74.8_{\pm 0.2}$	$65.6_{\pm 1.1}$	$38.6_{\pm 1.7}$	23M
V-REx	$72.7_{\pm 1.7}$	$60.8_{\pm 5.5}$	$24.5_{\pm 6.4}$	23M
CORAL	$73.6_{\pm 0.4}$	$62.8_{\pm 2.7}$	$37.9_{\pm 0.7}$	23M
Fish	$64.4_{\pm 2.0}$	$53.2_{\pm 4.5}$	$42.0_{\pm 0.8}$	23M
GroupDRO	$73.6_{\pm 2.1}$	$66.0_{\pm 3.8}$	$36.4_{\pm 0.9}$	23M
JTT	$74.4_{\pm 0.6}$	$64.6_{\pm 2.3}$	$33.8_{\pm 0.0}$	23M
DM-ADA	$74.0_{\pm 0.8}$	$65.7_{\pm 1.4}$	$35.8_{\pm 1.0}$	23M
LISA	$70.0_{\pm 0.7}$	$59.8_{\pm 2.3}$	$38.0_{\pm 1.3}$	23M
DISC	$75.5_{\pm 1.1}$	$73.5_{\pm 1.4}$	$55.1_{\pm 2.3}$	23M
CPT	$\mathbf{79.8}_{\pm 2.6}$	$\mathbf{77.0}_{\pm 0.8}$	$\mathbf{64.9}_{\pm 1.2}$	**47k**

MetaShift and ISIC. Results on those datasets are depicted in Table 3, along with the number of parameters required for each baseline. We mostly pick those that utilize group information during training for better benchmarking the efficiency of our proposed framework. In summary, the performance of CPT exceeds other methods by large margins, especially ≈ 10 AUROC score on ISIC. On MetaShift, it is also the best method in terms of average and worst-group accuracy, improving the performance by approximately 4%.

5.5. Ablation studies

In this section, we provide different ablation studies to see how each component in our proposed method individually contributes to the overall improvement of CPT.

Optimization and Backbones. In Table 4, $\text{CPT}_{\text{w/o entropy}}$ eliminates the entropy objective and uses the average gradient $1/K \sum \nabla \ell_i(\theta)$ to update the model. While this strategy can improve mean accuracy among groups, compared to ERM, the performance has a significant drop compared to our optimization procedure. This conclusion suggests that conducting simple gradient averaging when there are still large differences among group losses could cause severe bias in training. On the other hand, there is no significant difference between the performance of fully fine-tuning ResNet50 and prompt tuning on ViT-B/16, which indicates that our method does not benefit much from more powerful backbones. Instead, we remind readers that prompting in CPT is introduced to scale up our proposed algorithm. Thus, even applying our proposed balancing method alone can boost the performance of ResNet to exceed GroupDRO by large margins on WaterBirds (5.2%) and CelebA (5.1%).

Table 4: Ablation results for different architectures and optimization procedures.

Dataset	Method	Average	Worst	# Params
MetaShift	ERM	$72.9_{\pm 1.4}$	$62.1_{\pm 4.8}$	23M
	GroupDRO	$73.6_{\pm 2.1}$	$66.0_{\pm 3.8}$	23M
	DISC	$75.5_{\pm 1.1}$	$73.5_{\pm 1.4}$	23M
	$\text{CPT}_{\text{w/o entropy}}$ (ResNet50)	$76.1_{\pm 0.7}$	$62.8_{\pm 1.8}$	23M
	CPT (ResNet50)	$78.2_{\pm 2.5}$	$76.9_{\pm 2.0}$	23M
	ERM (prompt-tuning)	$78.4_{\pm 1.4}$	$65.0_{\pm 2.4}$	**47**
	$\text{CPT}_{\text{w/o entropy}}$	$77.3_{\pm 0.8}$	$74.7_{\pm 2.5}$	**47k**
	CPT	$\mathbf{79.8}_{\pm 2.6}$	$\mathbf{77.0}_{\pm 0.8}$	**47k**
Waterbirds	ERM	$97.1_{\pm 0.1}$	$70.0_{\pm 2.3}$	23M
	GroupDRO	$93.2_{\pm 0.5}$	$86.7_{\pm 0.6}$	23M
	CPT (ResNet50)	$90.9_{\pm 0.2}$	$91.9_{\pm 0.3}$	23M
	ERM (prompt-tuning)	$97.5_{\pm 0.2}$	$64.4_{\pm 2.6}$	**94k**
	CPT	$96.3_{\pm 0.2}$	$\mathbf{93.5}_{\pm 0.4}$	**94k**
CelebA	ERM	$94.8_{\pm 0.2}$	$45.0_{\pm 1.5}$	23M
	GroupDRO	$92.9_{\pm 0.3}$	$86.3_{\pm 1.1}$	23M
	CPT (ResNet50)	$92.7_{\pm 0.3}$	$91.4_{\pm 0.8}$	23M
	ERM (prompt-tuning)	$95.6_{\pm 0.3}$	$58.3_{\pm 1.9}$	**94k**
	CPT	$93.2_{\pm 0.3}$	$\mathbf{92.0}_{\pm 0.3}$	**94k**

Gap among group performance. Apart from Figure 1, we visualize the performance of each comparative method at its early stopping epoch and the point when it actually obtains the highest score on the minority group in Figure 4. Those points collapse in the case of ERM since it does not have any balancing mechanism. Unsurprisingly, the gap between minority and majority performance of ERM is largest ($> 10\%$) while this figure decreases for GroupDRO and CPT. It is true that GroupDRO can obtain a relatively high score on the minority group early in training, even on par with CPT. However, its majority group performance at this point has not converged yet ($\approx 90.5\%$), and when converged, its performance on the minority group declines considerably ($91\% \rightarrow 88\%$). By contrast, the is only a small difference in terms of minority performance for CPT at those points, allowing to have equally good performance across groups through training.

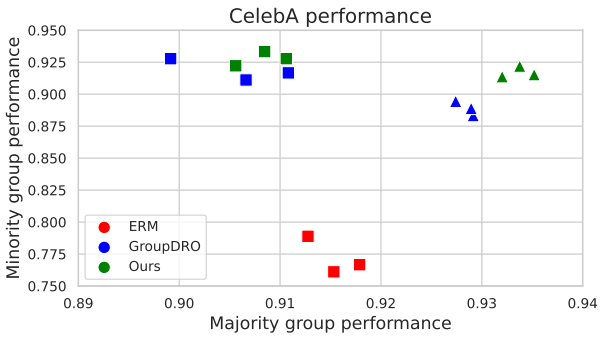


Figure 4: Performance of ResNet50 at: the last checkpoint used for evaluation (highest worst group accuracy on validation set), denoted by $\triangle$, and the checkpoint where the performance on the minority group is highest, denoted by $\square$. Results are obtained on three random seeds.

Effect of controlling vector Table 5 presents the distribution of the CelebA dataset and the performance of GroupDRO and our proposed method with different values of the controlling vectors. We can see that setting c equal to the default value $(1, 1, 1, 1)$ can perfectly balance the accuracy score across groups while GroupDRO encounters difficulty in fitting the minority one. Furthermore, by varying the value of c, we could observe different behavior of the model in each group. For example, the accuracy scores on two smaller groups (2&3) increase significantly, as a result of the loss magnitudes decrease when we put more weight on penalizing those two. With a tiny sacrifice on the performance of two majority groups, the model could gain considerable improvement in minority groups and the overall performance.

Table 5: Performance of the model after training for one epoch. Here, we simply adjust the optimization in favor of two minority groups to observe their improvement.

Group g (size)	**0** (44%)	**1** (41%)	**2** (14%)	**3** (1%)	Worst	Avg	Mean
GroupDRO	92.0	92.5	91.5	86.1	86.1	92.08	89.4
CPT $c = (1, 1, 1, 1)$	**92.9**	**93.4**	92.8	91.7	**91.7**	**93.1**	92.7
CPT $c = (1, 1, 1, 2)$	92.0	91.1	93.1	**95.6**	91.1	91.9	93.0
CPT $c = (1, 1, 2, 1)$	91.2	91.8	**95.2**	95.0	91.2	92.0	**93.3**

Saliency maps. We plot the saliency maps (Selvaraju et al., 2017) produced when predicting the target attribute of in-domain images in Figure 5 and out-of-domain images in Figure 6. Pixels highlighted by warmer colors indicate higher contributions to the model predictions. Our proposed method can learn causal features rather than focusing on spurious background features when making predictions. Hence, balancing the learning among groups not only de-bias the classifier in-domain but also robustifies it in tackling distributional shifts.

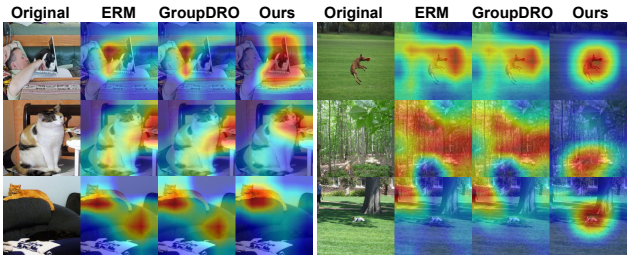


Figure 5: GradCAM of ResNet50 on in-domain samples.

Running time Figure 7 summarizes the total runtime and the time for solving the linear programming problem on each epoch. The introduction of the optimization (7) in the main paper leads to a small overhead in terms of training time. However, this small sacrifice significantly boosts the performance of different backbones, especially for CLIP and prompt-based models. For example, prompt-tuning on CLIP variants takes only 0.3 second per epoch since it does not have to update the entire network or even the classification head.

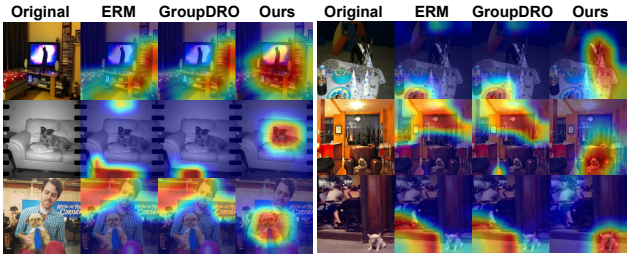


Figure 6: GradCAM of ResNet50 on OOD samples.

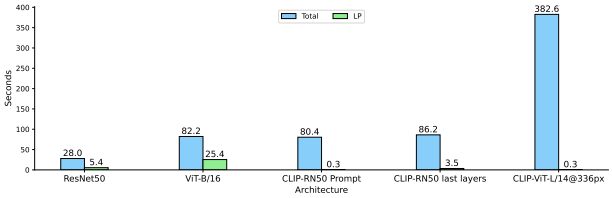


Figure 7: Total running time and the time for solving the linear programming (LP) problem on epoch with Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz and NVIDIA A100-SXM4-80GB GPU. Results are averaged over 5 epochs.

6. Limitations

We discuss the limitations of our approach here. While we scale our method when the number of controlling vectors grows via prompt-tuning, we indeed still have to train different prompt sets separately and individually for predetermined values of c (Table 5). Furthermore, in cases where those models could not perform as expected, one might consider incorporating another controlling vector which would require retraining another prompt set from scratch.

To address this limitation, one promising approach is employing hypernetworks (Qu et al., 2022; Do et al., 2023; Jiang et al., 2023), which take controlling vectors as input to generate corresponding prompt sets. This enables us to train the hypernetwork in an end-to-end fashion without predefining c and maintaining separate prompt sets. Despite this potential limitation, our method is generally tractable and readily applicable.

7. Conclusion

In this paper, we propose Controllable Prompt Tuning, a novel training method that not only improves the performance across groups but also allows us to control the trade-off between them. Our proposed method can effectively de-bias the training of different types of backbones and achieves state-of-the-art performance on benchmark datasets.

Acknowledgements

We thank Chau Pham, Polina Kirichenko, and Shikai Qiu for their helpful discussions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., and Brundage, M. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Ahmed, F., Bengio, Y., Van Seijen, H., and Courville, A. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2020.

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Andersen, M., Dahl, J., and Vandenberghe, L. Cvxopt: Convex optimization. *Astrophysics Source Code Library*, pp. ascl–2008, 2020.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., and Kurakin, A. Adamatch: A unified approach to semi-supervised learning and domain adaptation. In *International Conference on Learning Representations*, 2021.

Bissoto, A., Valle, E., and Avila, S. Debiasing skin lesion datasets and models? not so fast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 740–741, 2020.

Bradley, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Cai, T., Gao, R., Lee, J., and Lei, Q. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pp. 1170–1182. PMLR, 2021.

Chen, C.-F. R., Fan, Q., and Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.

Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., and Luo, P. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022a.

Chen, Y., Rosenfeld, E., Sellke, M., Ma, T., and Risteski, A. Iterative feature matching: Toward provable domain generalization with logarithmic environments. *Advances in Neural Information Processing Systems*, 35:1725–1736, 2022b.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

Cortes, C., Mohri, M., and Muñoz Medina, A. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178, 2015.

Dehdashtian, Sepehr Wang, L. and Boddeti, V. Fairerclip: Debiasing zero-shot predictions of clip in rkhss. In *International Conference on Learning Representations*, 2024.

Deng, Y., Yang, Y., Mirzasoleiman, B., and Gu, Q. Robust learning with progressive data expansion against spurious correlation. *Advances in neural information processing systems*, 2023.

Désidéri, J.-A. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

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.

Diamond, S. and Boyd, S. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

Do, G., Le, K., Pham, Q., Nguyen, T., Doan, T.-N., Nguyen, T.-B., Liu, C., Ramasam, S., Li, X., and Hoi, S. Hyperrouter: Towards efficient training and inference of sparse mixture of experts via hypernetwork. In *The 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 Main*, pp. 1–12, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., and Li, G. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022.

Edalati, A., Tahaei, M., Kobyzev, I., Nia, V. P., Clark, J. J., and Rezagholizadeh, M. Krona: Parameter efficient tuning with kronecker adapter. *arXiv preprint arXiv:2212.10650*, 2022.

Fan, A., Grave, E., and Joulin, A. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2019.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Gheini, M., Ren, X., and May, J. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1754–1765, 2021.

Ghosal, S. S. and Li, Y. Are vision transformers robust to spurious correlations? *International Journal of Computer Vision*, pp. 1–21, 2023.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Guo, D., Rush, A., and Kim, Y. Parameter-efficient transfer learning with diff pruning. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

He, H., Cai, J., Zhang, J., Tao, D., and Zhuang, B. Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11825–11835, 2023.

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., and Neubig, G. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Huang, Q., Dong, X., Chen, D., Zhang, W., Wang, F., Hua, G., and Yu, N. Diversity-aware meta visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10878–10887, 2023.

Jiang, G., JIANG, C., Xue, S., Zhang, J. Y., ZHOU, J., Lian, D., and Wei, Y. Towards anytime fine-tuning: Continually pre-trained language models with hypernetwork prompts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Kim, N., Kang, J., Ahn, S., Ok, J., and Kwak, S. Removing multiple biases through the lens of multi-task learning. *International Conference on Machine Learning*, 2023.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2022.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Kumar, A., Ma, T., Liang, P., and Raghunathan, A. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *Uncertainty in Artificial Intelligence*, pp. 1041–1051. PMLR, 2022.

Lei, Q., Hu, W., and Lee, J. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pp. 6164–6174. PMLR, 2021.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021.

Li, B., Hu, Y., Nie, X., Han, C., Jiang, X., Guo, T., and Liu, L. Dropkey for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22700–22709, 2023.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics)*, pp. 4582–4597, 2021.

Lialin, V., Deshpande, V., and Rumshisky, A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

Liang, J., Gong, K., Li, S., Liu, C. H., Li, H., Liu, D., Wang, G., et al. Pareto domain adaptation. *Advances in Neural Information Processing Systems*, 34:12917–12929, 2021.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2021.

Lin, X., Zhen, H.-L., Li, Z., Zhang, Q.-F., and Kwong, S. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

Liu, B., Liu, X., Jin, X., Stone, P., and Liu, Q. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021a.

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, 2021b.

Liu, L., Li, Y., Kuang, Z., Xue, J.-H., Chen, Y., Yang, W., Liao, Q., and Zhang, W. Towards impartial multi-task learning. In *International Conference on Learning Representations*, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Mahapatra, D. and Rajan, V. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*, pp. 6597–6607. PMLR, 2020.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 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.

Oh, D., Lee, D., Byun, J., and Shin, B. Improving group robustness under noisy labels using predictive uncertainty. *arXiv preprint arXiv:2212.07026*, 2022.

Panahi, A., Saeedi, S., and Arodz, T. Shapeshifter: a parameter-efficient transformer using factorized reshaped matrices. *Advances in Neural Information Processing Systems*, 34:1337–1350, 2021.

Phan, H., Tran, L., Tran, N. N., Ho, N., Phung, D., and Le, T. Improving multi-task learning via seeking task-based flat regions. *arXiv preprint arXiv:2211.13723*, 2022a.

Phan, H., Tran, N., Le, T., Tran, T., Ho, N., and Phung, D. Stochastic multiple target sampling gradient descent. *Advances in neural information processing systems*, 35: 22643–22655, 2022b.

Phan, H., Le, T., Phung, T., Bui, A. T., Ho, N., and Phung, D. Global-local regularization via distributional robustness. In *International Conference on Artificial Intelligence and Statistics*, pp. 7644–7664. PMLR, 2023.

Piratla, V., Netrapalli, P., and Sarawagi, S. Focus on the common good: Group distributional robustness follows. In *International Conference on Learning Representations*, 2021.

Qiu, S., Potapczynski, A., Izmailov, P., and Wilson, A. G. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*. PMLR, 2023.

Qu, J., Faney, T., Wang, Z., Gallinari, P., Yousef, S., and de Hemptinne, J.-C. Hmoe: hypernetwork-based mixture of experts for domain generalization. *arXiv preprint arXiv:2211.08253*, 2022.

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.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

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, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shi, Y., Seely, J., Torr, P., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2021.

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.

Sun, B., Feng, J., and Saenko, K. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pp. 153–171, 2017.

Sung, Y.-L., Cho, J., and Bansal, M. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2022.

Taghanaki, S. A., Choi, K., Khasahmadi, A. H., and Goyal, A. Robust representation learning via perceptual similarity metrics. In *International Conference on Machine Learning*, pp. 10043–10053. PMLR, 2021.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tran, N. N., Duong, S., Phan, H., Pham, T., Phung, D., and Le, T. Sharpness & shift-aware self-supervised learning. *arXiv preprint arXiv:2305.10252*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Wang, J., Liu, Y., and Wang, X. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1995–2008, 2021.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022.

Wu, C.-E., Tian, Y., Yu, H., Wang, H., Morgado, P., Hu, Y. H., and Yang, L. Why is prompt tuning for vision-language models robust to noisy labels? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15488–15497, 2023a.

Wu, S., Yuksekgonul, M., Zhang, L., and Zou, J. Discover and cure: Concept-aware mitigation of spurious correlation. *International Conference on Machine Learning*, 2023b.

Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., and Zhang, W. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6502–6509, 2020.

Yang, S., Dong, Y., Ward, R., Dhillon, I. S., Sanghavi, S., and Lei, Q. Sample efficiency of data augmentation consistency regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3825–3853. PMLR, 2023a.

Yang, Y., Nushi, B., Palangi, H., and Mirzasoleiman, B. Mitigating spurious correlations in multi-modal models during fine-tuning. *International Conference on Machine Learning*, 2023b.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836, 2020.

Zaken, E. B., Goldberg, Y., and Ravfogel, S. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, 2022.

Zhai, R., Dan, C., Kolter, J. Z., and Ravikumar, P. K. Understanding why generalized reweighting does not improve over erm. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhang, M. and Ré, C. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Re, C. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, pp. 26484–26516. PMLR, 2022.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023a.

Zhu, J., Lai, S., Chen, X., Wang, D., and Lu, H. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9516–9526, 2023b.

Zitzler, E. and Thiele, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271, 1999.

Supplementary Materials for "**Controllable Prompt Tuning for Balancing Group Distributional Robustness**"

Due to space constraints, some details were omitted from the main paper. We therefore include proof for Theorem 4.1 in the main paper (Appendix A) and more detailed experimental results (Appendix B) in this supplementary material.

A. Proof of Theorem 4.1

Theorem 4.1. *Assume that the loss function ℓ is differentiable up to the first order with respect to θ, then following*

$$d_{\text{ent}} := \sum_{i=1}^{K} \nabla \ell_i(\theta) \left[p_i \log\left(p_i\right) - p_i \sum_{j=1}^{K} \log\left(p_j\right) p_j \right]$$

where $p_i = \frac{e^{\ell_i(\theta)}}{\sum_{j=1}^{K} e^{\ell_j(\theta)}}$, maximizes the objective $\mathcal{L}_{ent}(\theta)$:

$$\mathcal{L}_{ent}(\theta) = H\big(\text{softmax}\big(\vec{\mathcal{L}}(\theta)\big)\big) = H\Big(\text{softmax}\big(\big\{\mathbb{E}_{P_{g^k}}\left[\ell\left(f_\theta(x), y\right)\right]\big\}_{k=1}^{K}\big)\Big)$$

Proof: For brevity of notations, we denote $\mathbb{E}_{P_{g^i}}\left[\ell\left(f_\theta(x), y\right)\right]$ as $\ell_i(\theta)$ and $p_i = \frac{e^{\ell_i(\theta)}}{\sum_{j=1}^{K} e^{\ell_j(\theta)}}$. The derivative of $\mathcal{L}_{ent}$ with respect to θ is computed as:

$$\begin{aligned}
\nabla \mathcal{L}_{ent}(\theta) &= -\sum_{i=1}^{K} \left\{\nabla p_i [\log(p_i) + 1]\right\} = -\sum_{i=1}^{K} \left\{ [\log(p_i) + 1] \nabla \frac{e^{\ell_i(\theta)}}{\sum_{j=1}^{K} e^{\ell_j(\theta)}} \right\} \\
&= -\sum_{i=1}^{K} \left\{ [\log(p_i) + 1] \frac{\sum_{j=1}^{K} e^{\ell_j(\theta)} e^{\ell_i(\theta)} \nabla \ell_i(\theta) - e^{\ell_i(\theta)} \sum_{j=1}^{K} e^{\ell_j(\theta)} \nabla \ell_j(\theta)}{(\sum_{j=1}^{K} e^{\ell_j(\theta)})^2} \right\} \\
&= -\frac{1}{(\sum_{j=1}^{K} e^{\ell_j(\theta)})^2} \sum_{i=1}^{K} \left\{ [\log(p_i) + 1] e^{\ell_i(\theta)} \sum_{j=1}^{K} e^{\ell_j(\theta)} (\nabla \ell_i(\theta) - \nabla \ell_j(\theta)) \right\} \\
&= -\frac{1}{(\sum_{j=1}^{K} e^{\ell_j(\theta)})^2} \sum_{i=1}^{K} \Big[\log(p_i) e^{\ell_i(\theta)} \sum_{j=1}^{K} e^{\ell_j(\theta)} (\nabla \ell_i(\theta) - \nabla \ell_j(\theta)) \Big] \\
&= -\frac{1}{(\sum_{j=1}^{K} e^{\ell_j(\theta)})^2} \sum_{i=1}^{K} \nabla \ell_i(\theta) \Big[\log(p_i) e^{\ell_i(\theta)} \sum_{j=1}^{K} e^{\ell_j(\theta)} - e^{\ell_i(\theta)} \sum_{j=1}^{K} \log(p_j) e^{\ell_j(\theta)} \Big] \\
&= \sum_{i=1}^{K} \nabla \ell_i(\theta) [p_i \sum_{j=1}^{K} \log(p_j) p_j - p_i \log(p_i)]
\end{aligned} \tag{8}$$

We conclude the proof. It is noteworthy that when the loss function ℓ is positive, one can also simplify the computation by minimizing the following balancing function:

$$\mathcal{L}'_{ent}(\theta) := H\Big(\Big\{ \frac{\mathbb{E}_{P_{g^i}}\left[\ell\left(f_\theta(x), y\right)\right]}{\sum_{j=1}^{K} \mathbb{E}_{P_{g^j}}\left[\ell\left(f_\theta(x), y\right)\right]} \Big\}_{i=1}^{K} \Big)$$

Eliminating the exponential term and denoting $q_i = \frac{\ell_i(\theta)}{\sum_{j=1}^K \ell_j(\theta)}$, the gradient direction of the entropy term is given by:

$$\begin{aligned}\nabla \mathcal{L}'_{ent}(\theta) &= \sum_{i=1}^{K} \left\{ \nabla q_i [\log(q_i) + 1] \right\} \\ &= \sum_{i=1}^{K} \left\{ \frac{[\sum_{j=1}^K \ell_j(\theta)] \nabla \ell_i(\theta) - \ell_i(\theta)[\sum_{j=1}^K \nabla \ell_j(\theta)]}{[\sum_{j=1}^K \ell_j(\theta)]^2} [\log(q_i) + 1] \right\} \\ &\propto \sum_{i=1}^{K} \nabla \ell_i(\theta) \Big\{ \log(q_i) \big[\sum_{j=1}^{K} \ell_j(\theta) \big] - \sum_{j=1}^{K} \log(q_j) \ell_j(\theta) \Big\}\end{aligned}$$

B. Experiment details and additional empirical results

In this section, we first describe the datasets and models used in Appendix B.1 along with the detailed training configuration for our method, then provide additional results for the experiment in Appendix B.2. Example images and data statistics are presented in Table 7, 8, 9 and 10.

B.1. Implementation details

Hyper-parameter We mainly examine the proposed method on ResNet50 (He et al., 2016), ViT B/16 (Dosovitskiy et al., 2020) and different variants of OpenAI's CLIP[1] model. The hyperparameters selected for each experiment are given in Table 6. Unless stated otherwise, results of baselines are taken from original papers and (Deng et al., 2023; Wu et al., 2023b; Piratla et al., 2021; Zhang et al., 2022; Kim et al., 2023), which provide standard evaluation protocols for different backbones and datasets. Performance of ERM and GroupDRO (Sagawa et al., 2019) at different amount of training data and their training curves are obtained from the released codebase of GroupDRO[2] using their provided commands.

Dataset	Architecture	Learning Rate	Weight Decay	Batch Size	# Epochs	Prompt length
Waterbirds	ViT B/16	0.1	0.001	64	100	10
Waterbirds	Last layers CLIP-RN50	0.003	0.01	32	100	-
Waterbirds	Prompt CLIP-RN50	0.03	0.01	32	100	16
Waterbirds	Last layer CLIP-ViT-L/14	0.01	0.01	32	100	-
Waterbirds	Prompt CLIP-ViT-L/14	0.001	0.01	32	100	16
CelebA	ViT B/16	0.01	0.001	128	100	10
MetaShift	ResNet50	0.003	0.01	16	100	-
MetaShift	ViT B/16	0.01	0.1	16	100	5
ISIC	ViT B/16	0.003	0.01	16	100	5

Table 6: Hyperparameter for different experiments throughout our paper. We report the hyper-parameters selected for our proposed method after performing grid-search.

Classifier design. While we train a classification head from scratch for ResNet50 or ViT, CLIP allows us to utilize its meaningful multimodality embeddings as a zero-shot classifier. In particular, given an image embedding $\mathbf{e}_i$ and a text class embedding $\mathbf{e}_t^k$ for some class index $k \in \{1, 2, \ldots, \mathbf{N_c}\}$, where $\mathbf{N}_c$ is the number of categories. The probability of the given image belonging to k-th class is computed as:

$$p(y = k \mid \mathbf{e}_i) = \frac{\exp\left(< \mathbf{e}_i, \mathbf{e}_t^k > / \tau \right)}{\sum_{j=1}^{\mathbf{N}_c} \exp\left(< \mathbf{e}_i, \mathbf{e}_t^j > / \tau \right)}$$

where $< \cdot, \cdot >$ denotes the cosine similarity and τ is the temperature parameter.

Dataset statistics. We evaluate CPT on Waterbirds (Sagawa et al., 2019) (Table 7), CelebA (Liu et al., 2015) (Table 8), MetaShift (Liang & Zou, 2021) (Table 9) and ISIC (Codella et al., 2019) (Table 10), following previous work (Sagawa et al., 2019; Deng et al., 2023; Wu et al., 2023b). While the objective of those image classification tasks is predicting the

[1] https://github.com/openai/CLIP
[2] https://github.com/kohpangwei/group_DRO

categories of input images, those target attributes are often correlated well with spurious features (please refer to each table for detailed descriptions of spurious features and group partitions for each dataset).

Image				
Group	0	1	2	3
Label	0 (landbird)	0 (landbird)	1 (waterbird)	1 (waterbird)
Spurious feature	0 (land)	1 (water)	0 (land)	1 (water)
Description	landbird on land	landbird on water	waterbird on land	waterbird on water
# Train data	$3,498(73\%)$	$184(4\%)$	$56(1\%)$	$1,057(22\%)$
# Val data	467	466	133	133
# Test data	$2,255$	$2,255$	642	642

Table 7: Example images of Waterbirds (Sagawa et al., 2019).

Image				
Group	0	1	2	3
Label	0 (non-blond)	0 (non-blond)	1 (blond)	1 (blond)
Spurious feature	0 (woman)	1 (man)	0 (woman)	1 (man)
Description	non-blond woman	non-blond man	blond woman	blond man
# Train data	$71,629(44\%)$	$66,874(41\%)$	$22,880(14\%)$	$1,387(1\%)$
# Val data	$8,535$	$8,276$	$2,874$	182
# Test data	$9,767$	$7,535$	$2,480$	180

Table 8: Example images of CelebA (Liu et al., 2015).

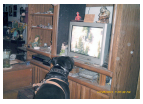
Image						
Group	0	1	2	3	4	5
Label	0 (cat)	0 (cat)	1(dog)	1 (dog)	0 (cat)	1 (dog)
Spurious feature	0 (sofa)	1 (bed)	2 (bench)	3 (bike)	4 (shelf)	4 (shelf)
# Train data	231	380	145	367	-	-
# Val data (OOD)	-	-	-	-	34	47
# Test data	-	-	-	-	201	259

Table 9: Example images of MetaShift (Liang & Zou, 2021).

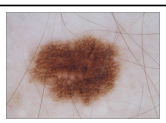
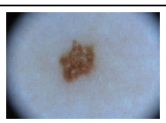
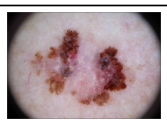
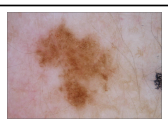
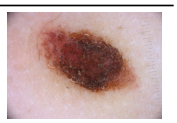
Image				...		
Label	0 (benign)	0 (benign)	1 (malignant)	...	0 (benign)	1 (malignant)
Spurious feature	hair	dark corner	gel bubbles	...	ink	ruler
# Train data:			1,826			
# Val data:			154			
# Test data:			618			

Table 10: Example images of ISIC (Codella et al., 2019).

B.2. Additional experiments

Prompt length Thanks to the design of the prompting component, the computational burden during training is modest: only a small portion of the model parameters are updated through back-propagation. Here, we vary the value of prompt length L to observe its effect on the model performance in Table 11. Results of ERM, GroupDRO, the second-best method in this dataset DFR (Kirichenko et al., 2022) and a recent baseline in this line of research PDE (Deng et al., 2023) are also included for a general comparison. While increasing the prompt length can improve the performance of the model, we find that just pretending one token per transformer layer ($L = 1$) can outperform the second-best method that requires $\times 2200$ more trainable parameters via better training. Last, it is noteworthy to highlight that $L = 16$ can help CPT match the average accuracy score of ERM in this scenario, which demonstrates the large modeling capacity of our proposed method.

Table 11: Ablation results for different prompt length N on ViT B/16 on Waterbirds.

N	1	2	4	8	10	16	ERM	GroupDRO	PDE	DFR
Worst	$92.9_{\pm 0.8}$	$93.3_{\pm 0.9}$	$93.5_{\pm 0.6}$	$93.8_{\pm 0.5}$	$93.5_{\pm 0.4}$	$\mathbf{93.9}_{\pm 0.7}$	$70.0_{\pm 2.3}$	$86.7_{\pm 0.6}$	$90.3_{\pm 0.3}$	$92.9_{\pm 0.2}$
Average	$95.8_{\pm 0.9}$	$95.9_{\pm 0.9}$	$96.4_{\pm 0.1}$	$96.5_{\pm 0.0}$	$96.3_{\pm 0.1}$	$\mathbf{97.1}_{\pm 0.3}$	$\mathbf{97.1}_{\pm 0.1}$	$93.2_{\pm 0.5}$	$92.4_{\pm 0.8}$	$94.2_{\pm 0.2}$
Mean	$94.9_{\pm 0.8}$	$95.2_{\pm 0.0}$	$95.6_{\pm 0.4}$	$\mathbf{96.0}_{\pm 0.8}$	$95.7_{\pm 0.2}$	$95.7_{\pm 0.3}$	N/A	N/A	N/A	N/A
# Params	10754	19970	38402	75266	93698	148994	23512130	23512130	23512130	23512130

Performance at different number of training samples. In order to comprehensively assess the performance of our proposed model, we conduct a series of experiments to analyze its behavior across varying sizes of training data. Those experiments were designed to investigate the impact of data size on the model performance of comparative methods. We keep group ratios preserved while subsampling a part of the training data. The x-axis of Figure 8 denotes the percentages of the Waterbirds and CelebA training set used in each sub-experiment. It can be seen that CPT is better than GroupDRO or ERM in lower data regimes. Remarkably, by using 20% and 10% data, our proposed method can still outperform other baselines when using the full training set.

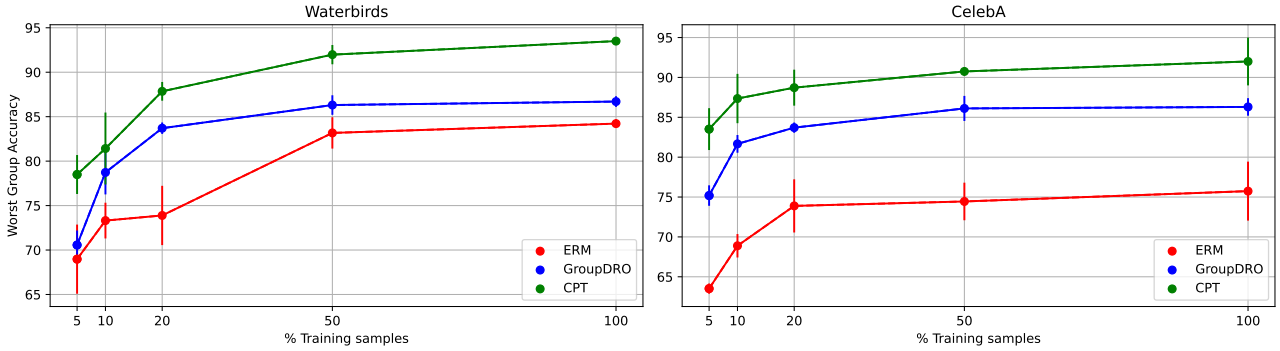


Figure 8: Data Efficiency on Waterbirds (left) and CelebA (right). Error bars correspond to standard deviations over three trials.

Similarly, we run the above experiment but replace the main backbone with CLIP-RN50 and plot the performance of different baselines. Note that we freeze both image and text encoders and employ prompt tuning on their text encoder only. The text prompt used for training CLIP models is:

"a type of bird, a photo of a waterbird/landbird"

While the pretrained CLIP model obtains 90.9% average accuracy on Waterbirds (Table 2 in the main paper), it severely suffers from unintended bias when attaining only 44.9% worst-group accuracy. This motivates the need for debiasing methods to remove such bias and mitigate the reliance on spurious features when making predictions of foundation models. Figure 9 shows that while using more data can help enhance the pre-trained CLIP model better for all methods, CPT obtains the best worst group accuracy for every amount of data and surpasses the current state-of-the-art method (Yang et al., 2023b).

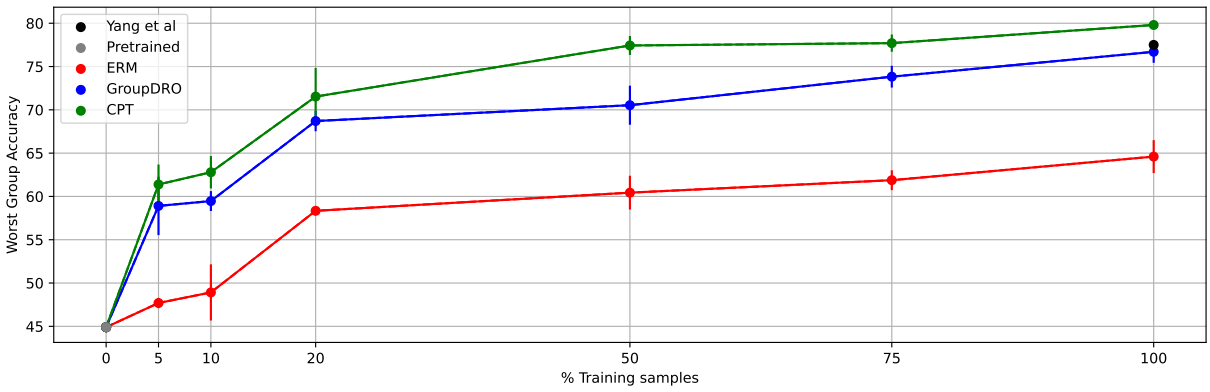


Figure 9: Data Efficiency of CLIP-RN50. CPT with 75% data can outperform the current state-of-the-art method for debiasing CLIP (Yang et al., 2023b) which uses $\times 1800$ trainable parameters.

Debiasing ViT. For the experiment of debiasing Vision Transformer, we compare CPT against different ViT and Big Transfer (BiT) models introduced in (Kolesnikov et al., 2020; Ghosal & Li, 2023). More detailed results are given in Table 12, from which we can see that CPT not only obtains the best worst-group accuracy score but also performs on par with the second-best method in terms of average accuracy while using much fewer parameters. Note that other models scale up the resolution to 384×384 while CPT keeps using the 224×224 resolution.

Table 12: Comparison of different backbones on Waterbirds

Model	Average Acc.	Worst-Group Acc.	# Params
BiT-M-R50x1	$92.05_{\pm 0.05}$	$75.10_{\pm 0.62}$	23.5M
BiT-M-R50x3	$94.90_{\pm 0.05}$	$80.51_{\pm 1.02}$	211M
BiT-M-R101x1	$94.05_{\pm 0.07}$	$77.50_{\pm 0.50}$	42.5M
ViT-Ti/16	$89.50_{\pm 0.05}$	$71.65_{\pm 0.16}$	5.6M
ViT-S/16	$96.30_{\pm 0.51}$	$85.45_{\pm 1.16}$	21.8M
ViT-B/16	$96.75_{\pm 0.05}$	$89.30_{\pm 1.95}$	86.1M
CPT	$96.33_{\pm 0.15}$	$\mathbf{93.51}_{\pm 0.36}$	**94**k

Performance on CLIP variants. Table 13 presents the scores of ERM, GroupDRO, and CPT on a wide range of CLIP backbones, including CLIP with ResNet101, ViT-B/16, ViT-B/32, ViT-L/14. Here we simply use the balanced version of CPT instead of hyperparameter tuning on c. Therefore, note that we do not try to beat GroupDRO on all metrics, but focus on investigating how our optimization procedure helps close the gap among groups. In summary, balancing the learning among groups results in large improvements in worst-group and mean accuracies for all backbones.

Table 13: Performance on other CLIP variants, where CPT has the lowest gap between worst and average accuracy.

Model	Last layers ResNet101			Prompting ResNet101			Last layer ViT-B/16			Prompting ViT-B/16		
# Params	13740544			8192			393216			8192		
Accuracy	Average	Worst	Mean	Average	Worst	Mean	Average	Worst	Mean	Average	Worst	Mean
Pre-trained	91.7	46.8	71.1	91.7	46.8	71.1	90.6	50.1	75.5	90.6	50.1	75.5
ERM	89.3	69.4	80.2	88.6	67.2	80.0	95.7	77.4	87.8	96.1	82.2	89.7
GroupDRO	88.5	74.9	82.1	86.3	74.0	80.7	91.4	88.5	90.2	93.3	85.5	90.4
CPT	87.7	**75.2**	**83.9**	83.1	**78.9**	**81.2**	91.4	**89.4**	**90.5**	92.2	**89.2**	**91.5**

Model	Last layer ViT-B/32			Prompting ViT-B/32			Last layer ViT-L/14			Prompting ViT-L/14		
# Params	393216			8192			786432			12288		
Accuracy	Average	Worst	Mean	Average	Worst	Mean	Average	Worst	Mean	Average	Worst	Mean
Pretrained	90.1	49.0	72.2	90.1	49.0	72.2	89.4	36.1	83.9	89.4	36.1	83.9
ERM	94.7	73.5	85.2	94.0	76.0	85.7	96.6	85.8	91.8	96.8	83.1	91.6
GroupDRO	90.7	81.6	86.8	88.3	82.3	85.9	95.8	88.9	92.8	95.5	86.3	91.5
CPT	87.4	**86.3**	**87.4**	86.8	**85.2**	**87.5**	94.8	**90.7**	**92.9**	94.9	**90.0**	**92.6**

Prompt-tuning with decoupled backbones. While the proposed method works well with CLIP, which has jointly-trained image and text encoders, we also examine it on decoupled backbones to show that the efficiency of CPT does not rely on powerful pretrained models. We thus accordingly expanded our experimental scope to include decoupled backbones and provide the results in Table 14. Since there is some difference between the original architecture of CLIP's image encoder and the ImageNet-pre-trained encoder, we need to train additional bottleneck layers to map both image and text inputs to a unified dimensional space.

Table 14: Performance of decoupled vision-language backbones on Waterbirds

Backbone	Method	Average	Worst	# Params
CLIP-RN50	Pre-trained	90.8	44.9	0
	ERM	85.7	64.6	8192
	GroupDRO	81.0	76.7	8192
	Ours	81.8	79.8	8192
CLIP & RN50 ImageNet	ERM	93.9	69.8	2105344
	GroupDRO	92.3	75.7	2105344
	Ours	92.4	79.3	2105344
CLIP-ViT-B/16	Pre-trained	90.6	50.1	0
	ERM	96.1	82.2	8192
	GroupDRO	93.3	85.5	8192
	Ours	92.2	89.2	8192
CLIP & ViT-B/16 ImageNet	ERM	75.5	47.9	401408
	GroupDRO	71.5	55.5	401408
	Ours	70.9	58.6	401408

Since there are mismatches between how the image and text encoders comprehend and represent information, the performance of those decoupled CLIP models drop significantly. However, in such cases, CPT still consistently exhibits superior worst-group accuracy compared to ERM or GroupDRO.

Label noise robustness. CPT is empowered by prompt-tuning, which has been shown to be more robust to noisy labels than full fine-tuning or linear probing, as discussed in (Wu et al., 2023a), underscoring its practical advantage in the outliers problem. To examine the robustness against noisy labels of different training methods, we conducted an additional experiment on the CLIP-RN50 and Waterbird dataset under different levels of label noise, ranging from 20% to 60%.

Table 15: Accuracy for models trained on Waterbirds with different levels of noisy labels.

	20%			**40%**			**60%**		
Method	Average	Mean	Worst	Average	Mean	Worst	Average	Mean	Worst
Last Layers (Yang et al., 2023b)	79.27 ± 2.74	75.85 ± 1.89	66.15 ± 2.31	67.16 ± 2.20	63.70 ± 2.32	57.84 ± 2.12	42.51 ± 1.49	45.59 ± 3.72	39.37 ± 1.50
Linear Probing	79.47 ± 2.48	75.76 ± 0.42	70.44 ± 2.67	67.05 ± 2.56	63.15 ± 0.96	57.06 ± 1.61	40.16 ± 1.48	40.79 ± 2.05	35.39 ± 2.08
Prompt-Tuning	77.66 ± 0.61	74.20 ± 1.43	70.35 ± 2.17	69.76 ± 2.89	64.71 ± 1.33	58.82 ± 1.67	50.00 ± 1.14	46.69 ± 1.60	42.76 ± 1.73
CPT	76.57 ± 1.65	75.38 ± 1.69	75.14 ± 2.51	67.67 ± 2.02	67.48 ± 0.43	61.83 ± 3.22	51.61 ± 1.06	49.91 ± 0.99	46.08 ± 1.28

As can be seen from Table 15, while those baselines achieve comparable performance at 20% noise rates, the accuracy scores of last layers tuning and linear probing drop significantly as the noise rate increases, compared to prompt tuning and CPT. In summary, CPT consistently achieves best worst-group accuracy in all settings and prompt. Interestingly, our optimization procedure helps improve prompt tuning on all metrics when the noise level is higher than 50%, which shows the benefit of effectively leveraging gradient information from multiple groups in deriving the updating direction.

Saliency maps. We here provide more GradCAM visual explanations for ERM, GroupDRO and CPT on Waterbirds in Figure 10. From those images, CPT shows its ability to identify causal features for making predictions, compared to ERM and GroupDRO. Hence, our balancing mechanism helps prevent the model from learning spurious features that occur frequently during training.

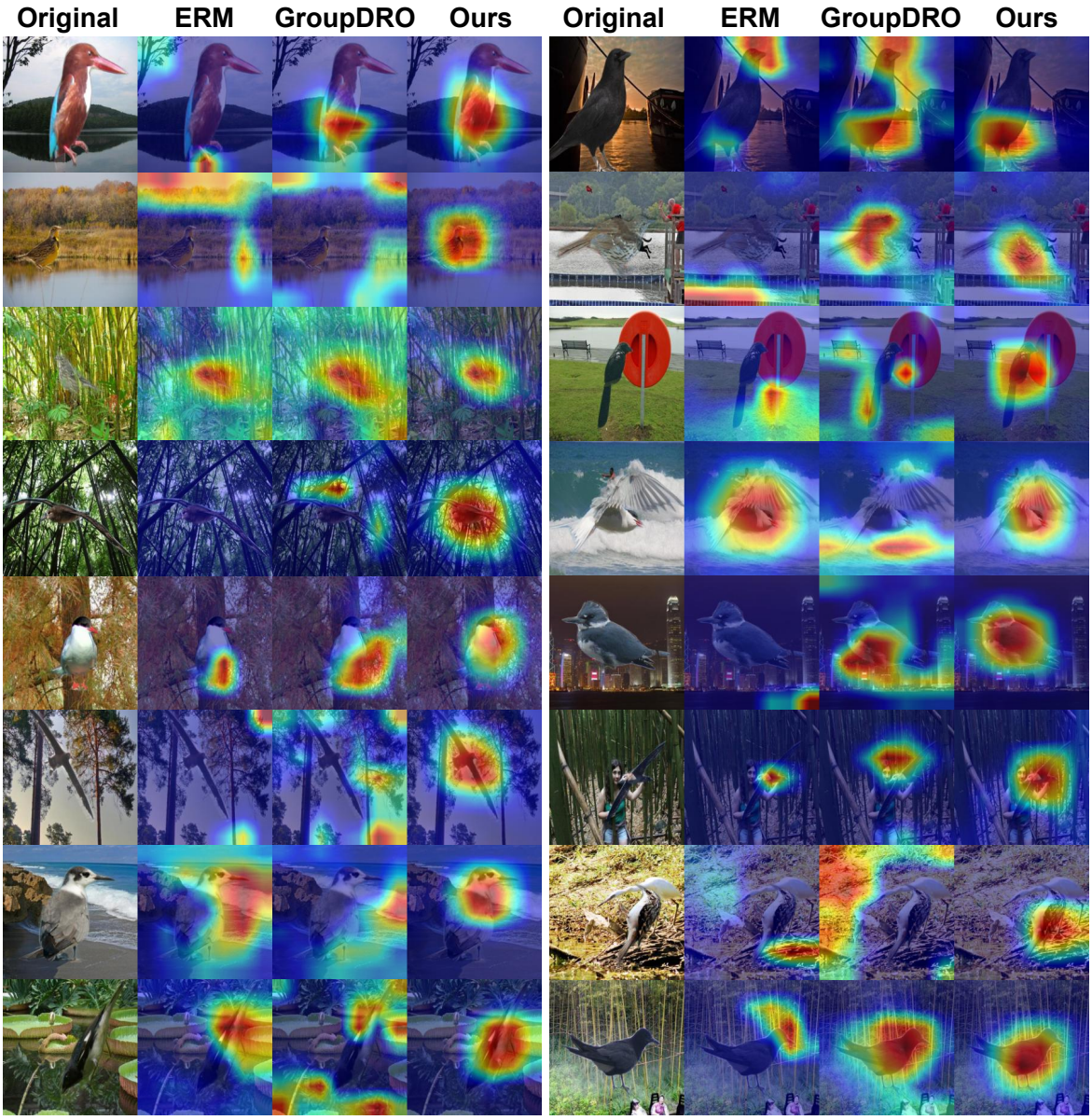


Figure 10: Waterbirds GradCAM.