Contents

1. Introduction **1**

2. Related Work **2**

3. Preliminaries **2**
3.1. Setup 2
3.2. Common Strategy for Improving Unsupervised Group Robustness 3

4. Methods **3**

5. Theoretical Analysis **4**
5.1. Removal of Class Proxies Amplifies Model Bias 4
5.2. Group Classification and Aggregation Mitigate Spurious Correlation 5

6. Experiments **5**
6.1. Datasets 5
6.2. Baselines 5
6.3. Training and Evaluation Details 6
6.4. Main Results 6
6.5. Further Analysis and Ablation Studies 7

7. Conclusion **8**

A. Proofs **12**
A.1. Proof of Proposition 1 12
A.2 Proof of Proposition 2 12

B. Additional Experimental Details **13**
B.1. Additional Implementation Details 13
B.2. Prompt Templates 13
B.3. Versatility of Fine-Tuning Paradigms 13
B.4. Noise Sensitivity of Pseudo-Labels 13
B.5. Comparison with Other Methods 13
B.6. Results without Tuning τ 14
B.7. Additional Dataset Details 14

C. Discussion of Limitation **14**

A. Proofs

A.1. Proof of Proposition 1

Restated Proposition (Proposition 1). *The weight of the spurious feature after projection is*

$$\gamma' = \gamma + \frac{\mathbf{r_s}^\top \mathbf{r}_{\mathbf{y}_o}}{\mathbf{r_s}^\top \mathbf{r_s}}. \tag{24}$$

Proof. We have $d \in \mathbb{R}^d$ core features $\mathbf{c}$ which determine the prediction target y and the spurious features s. Suppose we observe n features, stacked as $C = [\mathbf{c}_1, \cdots, \mathbf{c}_n] \in \mathbb{R}^{N \times d}$ and $\mathbf{s} = [s_1, \cdots, s_n] \in \mathbb{R}^N$. We have two regression scenarios:

- **Full model:** Linear regression on the core features $\mathbf{z}$ and the spurious feature s:

$$\mathbf{y} = C\boldsymbol{\beta} + \gamma \mathbf{s} + \boldsymbol{\varepsilon}, \tag{25}$$

  where $\boldsymbol{\beta}$ and γ are the weights associated with the core features $\mathbf{c}$ and the spurious feature s, respectively. $\boldsymbol{\varepsilon}$ is a noise term with an expected value of 0.
- **Projected model:** Applying the projection matrix Π to C to obtain $\tilde{C} = C\Pi$, followed by linear regression on $\tilde{\mathbf{c}}$ and s:

$$\mathbf{y} = \tilde{C}\tilde{\boldsymbol{\beta}} + \gamma' \mathbf{s} + \boldsymbol{\varepsilon}' \tag{26}$$

  where $\tilde{\boldsymbol{\beta}}$ and γ' are the weights for the projected core features $\tilde{\mathbf{c}}$ and the spurious feature s.

We define $M = I - \tilde{C}(\tilde{C}^\top \tilde{C})^{-1}\tilde{C}^\top$, $\mathbf{r_y} = M\mathbf{y}$ and $\mathbf{r_s} = M\mathbf{s}$. Applying M to both sides of Eq. (26), we obtain:

$$\begin{aligned} \mathbf{r_y} &= M\tilde{C}\tilde{\boldsymbol{\beta}} + \gamma' \mathbf{r_s} + M\boldsymbol{\varepsilon}' && (27)\\ &= \gamma' \mathbf{r_s} + M\boldsymbol{\varepsilon}' \qquad [M\tilde{C} = 0] && (28) \end{aligned}$$

The weight of spurious feature is derived as:

$$\gamma' = (\mathbf{r_s}^\top \mathbf{r_s})^{-1} \mathbf{r_s}^\top \mathbf{r_y}. \tag{29}$$

The core features C can be decomposed into the remaining part $\tilde{C}$ and the projected-out part C_o:

$$C = C\Pi + C(1 - \Pi) = \tilde{C} + C_o \tag{30}$$

Combine Eq. (30) and Eq. (25), we have:

$$\mathbf{y} = \tilde{C}\boldsymbol{\beta} + C_o\boldsymbol{\beta} + \gamma \mathbf{s} + \boldsymbol{\varepsilon} \tag{31}$$

Denote $\mathbf{y}_o = C_o\boldsymbol{\beta}$ is contribution of the projected-out core features and $\mathbf{r}_{\mathbf{y}_o} = M\mathbf{y}_o$, we can express $\mathbf{r_y}$ as:

$$\begin{aligned} \mathbf{r_y} &= M\mathbf{y} && (32)\\ &= M(\tilde{C}\boldsymbol{\beta} + C_o\boldsymbol{\beta} + \gamma \mathbf{s} + \boldsymbol{\varepsilon}) && (33)\\ &= \mathbf{r}_{\mathbf{y}_o} + \gamma \mathbf{r_s} + M\boldsymbol{\varepsilon} \qquad [M\tilde{C} = 0] && (34) \end{aligned}$$

Plugging into Eq. (29) and omitting the noise term (since $\mathbb{E}[\varepsilon \cdot s] = 0$), we have:

$$\begin{aligned} \gamma' &= (\mathbf{r_s}^\top \mathbf{r_s})^{-1} \mathbf{r_s}^\top (\mathbf{r_s}\gamma + \mathbf{r}_{\mathbf{y}_o}) && (35)\\ &= \gamma + (\mathbf{r_s}^\top \mathbf{r_s})^{-1} \mathbf{r_s}^\top \mathbf{r}_{\mathbf{y}_o} && (36) \end{aligned}$$

□

A.2. Proof of Proposition 2

Lemma 1. *Let $\hat{y} = \operatorname{argmax}_{y' \in \mathcal{Y}} f(\mathbf{x})_{y'}$ denote the prediction of f. The balanced group error (*BGE*) defined in Eq. (15) can be expressed as:*

$$\mathsf{BGE}(f) = \frac{1}{|\mathcal{G}|}\mathbb{E}_{\mathbf{x}}[\sum_{g \in \mathcal{G}} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \cdot \mathbb{P}(y \neq \hat{y}|\mathbf{x})]. \tag{37}$$

Proof.

$$\begin{aligned} \mathsf{BGE}(f) &= \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{E}_{\mathbf{x}|g}[y \neq \operatorname*{argmax}_{y' \in \mathcal{Y}} f(\mathbf{x})_{y'}] && (38)\\ &= \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \int_{\mathbf{x}} \mathbb{1}[y \neq \hat{y}] \mathbb{P}(\mathbf{x}|g) \mathrm{d}\mathbf{x} && (39)\\ &= \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \int_{\mathbf{x}} \mathbb{1}[y \neq \hat{y}] \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \mathbb{P}(\mathbf{x}) \mathrm{d}\mathbf{x} && (40)\\ &= \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{E}_{\mathbf{x}}[\frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \cdot \mathbb{1}[y \neq \hat{y}]] && (41)\\ &= \frac{1}{|\mathcal{G}|} \mathbb{E}_{\mathbf{x}}[\sum_{g \in \mathcal{G}} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \cdot \mathbb{1}[y \neq \hat{y}]] && (42) \end{aligned}$$

The expected value $\mathbb{E}_{\mathbf{x}}[\mathbb{1}[y \neq \hat{y}]]$ can be expressed as the joint expectation over $\mathbf{x}$ and y. By applying the law of total expectation, we rewrite Eq. (43) by conditioning on $\mathbf{x}$ in Eq. (44).

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}[\mathbb{1}[y \neq \hat{y}]] &= \mathbb{E}_{\mathbf{x},y}[\mathbb{1}[y \neq \hat{y}]] && (43)\\ &= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{y|\mathbf{x}}[\mathbb{1}[y = \hat{y}|\mathbf{x}]] && (44) \end{aligned}$$

Given $\mathbf{x}$, $\hat{y}$ is deterministic. Thus the inner expectation simplifies to the probability of $y = \hat{y}$ conditioned on $\mathbf{x}$:

$$\mathbb{E}_{y|\mathbf{x}}[\mathbb{1}[y = \hat{y}|\mathbf{x}] = \mathbb{P}(y = \hat{y}|\mathbf{x}) \tag{45}$$

Substituting back into Eq. (44), we have:

$$\mathbb{E}_{\mathbf{x}}[\mathbb{1}[y \neq \hat{y}]] = \mathbb{E}_{\mathbf{x}}[\mathbb{P}(y = \hat{y}|\mathbf{x}))] \tag{46}$$

Combining Eq. (46) with Eq. (42), we arrive at Eq. (37). □

Restated Proposition (Proposition 2). *Let $\mathcal{G}(y)$ denote the set of groups with class label y,* i.e., *$\mathcal{G}(y) := \{g =$*

$(y', a) \in \mathcal{G}|y' = y\}$. *Let* $\boldsymbol{\beta}$ *denote the group priors,* i.e., $\boldsymbol{\beta}_g = \mathbb{P}(g)$. *The prediction :*

$$\arg\max_{y\in\mathcal{Y}} f^*(\mathbf{x})_y = \arg\max_{y\in\mathcal{Y}} \sum_{g\in\mathcal{G}(y)} (h(\mathbf{x}) - \ln \boldsymbol{\beta})_g \quad (47)$$

is Bayes optimal for the problem in Eq. (15).

Proof. Using Lemma 1, to minimize the balanced group error, it is equivalent to minimize the term inside the expectation:

$$\sum_{g\in\mathcal{G}} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \cdot \mathbb{P}(y \neq \hat{y}|\mathbf{x}) \quad (48)$$

$$= \sum_{y\in\mathcal{Y}} [\sum_{g\in\mathcal{G}(y)} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \cdot (1 - \mathbb{P}(y = \hat{y}|\mathbf{x}))] \quad (49)$$

It is equivalent to maximize:

$$\sum_{y\in\mathcal{Y}} [\sum_{g\in\mathcal{G}(y)} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)}] \cdot \mathbb{P}(y = \hat{y}|\mathbf{x}). \quad (50)$$

Denote $a_y = \sum_{g\in\mathcal{G}(y)} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)}$ and $b_y = \mathbb{P}(y = \hat{y}|\mathbf{x})$. Since $\{a_y\}$ are fixed and b_y is a probability simplex, *i.e.*, $b_y > 0$ and $\sum_y b_y = 1$. We are equivalent to solving the following constrained optimization problem:

$$\max_{b_y} \sum_y a_y b_y, \text{ s.t. } b_y > 0, \sum_y b_y = 1 \quad (51)$$

It is straightforward to show that the optimal value is $\max_i a_i$, achieved when $i = \text{argmax}_i\, a_i$, with $b_i = 1$ and $b_j = 0$ for $j \neq i$. Substituting back the definitions of a and b. The solution is:

$$\mathbb{P}(y = \hat{y}|\mathbf{x}) = \begin{cases} 1, & \text{if } y = \arg\max_{y'} \sum_{g\in\mathcal{G}(y')} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \\ 0, & \text{otherwise} \end{cases} \quad (52)$$

It is equivalent to show that:

$$\mathbb{P}(\arg\max_{y'} \sum_{g\in\mathcal{G}(y')} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} = \hat{y}|\mathbf{x}) = 1 \quad (53)$$

Therefore, the *Bayes* optimal solution is:

$$\arg\max_{y\in\mathcal{Y}} f^*(\mathbf{x})_y = \arg\max_{y'} \sum_{g\in\mathcal{G}(y')} \frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \quad (54)$$

With the definition $\boldsymbol{\beta}_g = \mathbb{P}(g)$ and $\mathbb{P}(g|\mathbf{x}) \propto \exp(h(\mathbf{x})_g)$, we have:

$$\frac{\mathbb{P}(g|\mathbf{x})}{\mathbb{P}(g)} \propto \frac{\exp(h(\mathbf{x})_g)}{\exp(\ln \boldsymbol{\beta}_g)} \propto (h(\mathbf{x}) - \ln \boldsymbol{\beta})_g \quad (55)$$

Since proportional scaling does not change argmax results, combining Eq. (55) with Eq. (54), we obtain Eq. (47). □

B. Additional Experimental Details

B.1. Additional Implementation Details

For all methods evaluated in our experiments, we use the SGD optimizer with a weight decay of 5×10^{-5} and a momentum of 0.9. The initial learning rate is set to 0.0002 and decreases to 0 using cosine annealing. The models are trained for 100 epochs, with a warm-up learning rate of 10^{-5} applied during the first epoch to mitigate explosive gradients in the early training iterations. The batch size is set to 128 for most datasets, except for CelebA, where it is increased to 512 to accelerate training due to the dataset's relatively larger size. All classification heads including linear probing, prompt tuning and adapters are initialized with the zero-shot prompting. For all datasets except BAR, we evaluate the model on the validation set at the end of each epoch and select the one with the highest worst-group accuracy for final testing. For the BAR dataset, which lacks a validation set, we use the checkpoint from the last epoch for testing. The hyperparameter τ is searched within the range $[0.8, 0.9, 1.0, 1.1, 1.2]$. All experiments are conducted in a single NVIDIA A6000 GPU.

B.2. Prompt Templates

In Tab. 7, we present the prompt templates for zero-shot prompting and group-informed prompting for each dataset. Zero-shot prompting with class names is also used to construct the class proxy matrix Z in step 1 of our PPA. Tab. 8 lists the class names and group names for all datasets.

B.3. Versatility of Fine-Tuning Paradigms

In Tab. 9, we apply our PPA to other parameter-efficient fine-tuning paradigms using CLIP ViT-L/14 models. We observe consistent gains in worst group accuracies.

We further extend our method to train 2-Layer MLP after CLIP, with results in Tab. 10, showing that the 2-Layer MLP offers no significant gains over the linear layer.

B.4. Noise Sensitivity of Pseudo-Labels

To assess the noise sensitivity, we randomly select $p\%$ of the training samples and assign random values to subgroup labels within each class to introduce pseudo-label errors. The worst-group accuracies for varying p are shown in the figure below. Our results indicate that the proposed method maintains high WGA when label noise is below 10%, demonstrating its robustness under mild noise conditions.

B.5. Comparison with Other Methods

We compare our model with [16] using CLIP ResNet-50 in Tab. 11.

Table 7. Class prompt and group prompt templates.

Dataset	Class Prompt	Group Prompt
Waterbirds	*a type of bird, a photo of a* {*class*}.	*a type of bird, a photo of a* {*class*} *on* {*group*}.
CelebA	*a photo of a celebrity with* {*class*}.	*a photo of a celebrity, a* {*group*} *with* {*class*}.
MetaShift	*a photo of a* {*class*}.	*a photo of a* {*group*} {*class*}.
BAR	a photo of a person doing {class}.	N/A
Living-17	*a photo of a* {*class*}.	N/A

Table 8. Class and group names.

Dataset	Class Names	Group Names
Waterbirds	landbird, waterbird	land, water
CelebA	non-blond hair, blond hair	man, woman
MetaShift	dog, cat	outdoor, indoor
BAR	climbing, diving, fishing, pole vaulting, racing, throwing	N/A
Living-17	salamander, turtle, lizard, snake, spider, grouse, parrot, crab, dog, wolf, fox, cat, bear, beetle, butterfly, ape, monkey	N/A

Table 9. PPA consistently improves group robustness for across different efficient fine-tuning paradigms using CLIP ViT-L/14.

Method	Waterbirds		CelebA		MetaShift	
	WGA	Avg	WGA	Avg	WGA	Avg
Linear Probe + ERM	65.9	97.6	28.3	94.7	84.6	96.7
Linear Probe + PPA	87.2	94.6	90.4	91.0	94.8	96.8
CoOp + ERM	74.0	97.3	26.7	94.6	91.9	96.9
CoOp + PPA	87.4	94.1	85.6	88.3	93.7	96.4
Adapter + ERM	79.3	97.8	54.4	94.5	90.6	95.5
Adapter + PPA	83.3	95.8	88.3	91.7	92.3	96.4

Table 10. Results of other fine-Tuning paradigms.

	Waterbirds		CelebA		MetaShift	
	WGA	Avg	WGA	Avg	WGA	Avg
Linear Layer	84.3	88.3	91.1	92.1	90.8	94.7
2-Layer MLP	83.4	88.1	90.4	92.5	90.1	95.9
Full Fine-Tuning	83.7	89.8	91.8	93.2	89.5	95.2

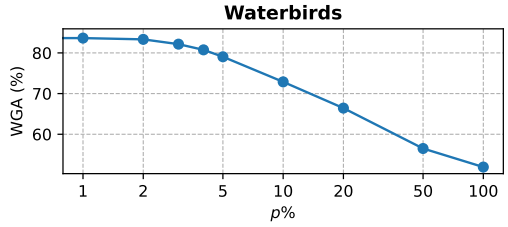

Figure 3. Results on noise sensitivity.

Table 11. Comparison with other methods.

Method	Waterbirds		CelebA	
	WGA	Avg	WGA	Avg
CLIP + B2T [2]	61.7	76.9	80.0	87.2
CLIP + PPA (ours)	84.3	88.3	91.1	92.1

B.6. Results without Tuning τ

The results with $\tau = 1$ are reported in Tab. 12. As expected, $\tau = 1$ still achieves SOTA.

Table 12. Results without Tuning τ.

	Waterbirds		CelebA		MetaShift	
	WGA	Avg	WGA	Avg	WGA	Avg
Optimal τ	84.3	88.3	91.1	92.1	90.8	94.7
$\tau = 1$	82.7	91.3	91.1	92.1	89.8	94.1

B.7. Additional Dataset Details

In this section, we show the statistics of all datasets used in our experiments in Tabs. 13 to 17 and illustrate some image samples in Figs. 4 to 7.

C. Discussion of Limitation

Our approach assumes that the CLIP text encoder can offer class proxies for downstream tasks. However, if the pre-trained knowledge diverges significantly from the downstream tasks, the effectiveness of our method may be lim-

Table 13. Statistics of Waterbirds.

	Train		Test	
	Water	Land	Water	Land
Waterbird	1057	56	642	642
Landbird	184	3498	2255	2255

Table 14. Statistics of CelebA.

	Train		Test	
	Female	Male	Female	Male
Blond	22880	1387	2480	180
Non-blond	71629	66874	9767	9767

Table 15. Statistics of MetaShift.

	Train		Test	
	Indoor	Outdoor	Indoor	Outdoor
Cat	630	153	345	65
Dog	402	635	191	273

Table 16. Statistics of Living-17.

	Train		Test	
	Majority	Minority	Majority	Minority
Group size	2340	117	100	100

Table 17. Statistics of BAR.

	Train		Test
	Majority	Minority	Minority
Climbing	326	5	100
Diving	520	8	151
Fishing	163	4	38
Racing	336	9	123
Throwing	137	3	82
Vaulting	279	7	124

ited. For instance, if the images are X-ray scans and the target is to predict a specific illness, the text encoder of the pre-trained model may lack relevant medical knowledge.

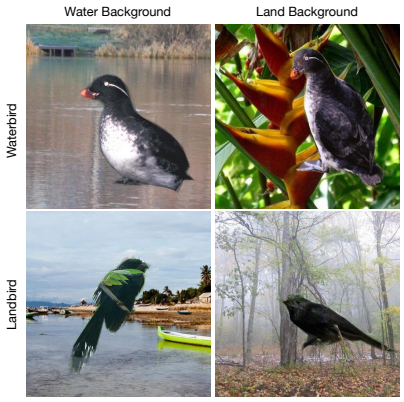

Figure 4. Image samples of Waterbirds.

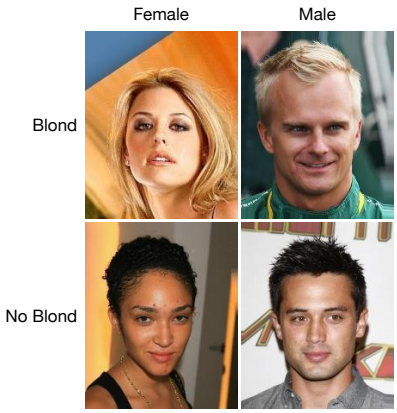

Figure 5. Image samples of CelebA.

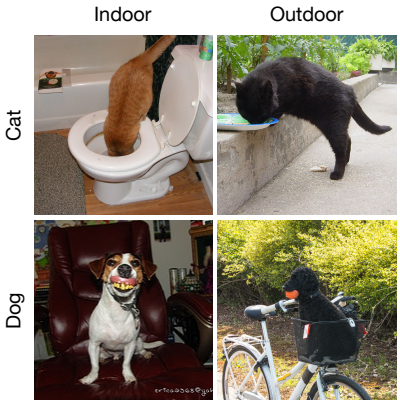

Figure 6. Image samples of MetaShift.

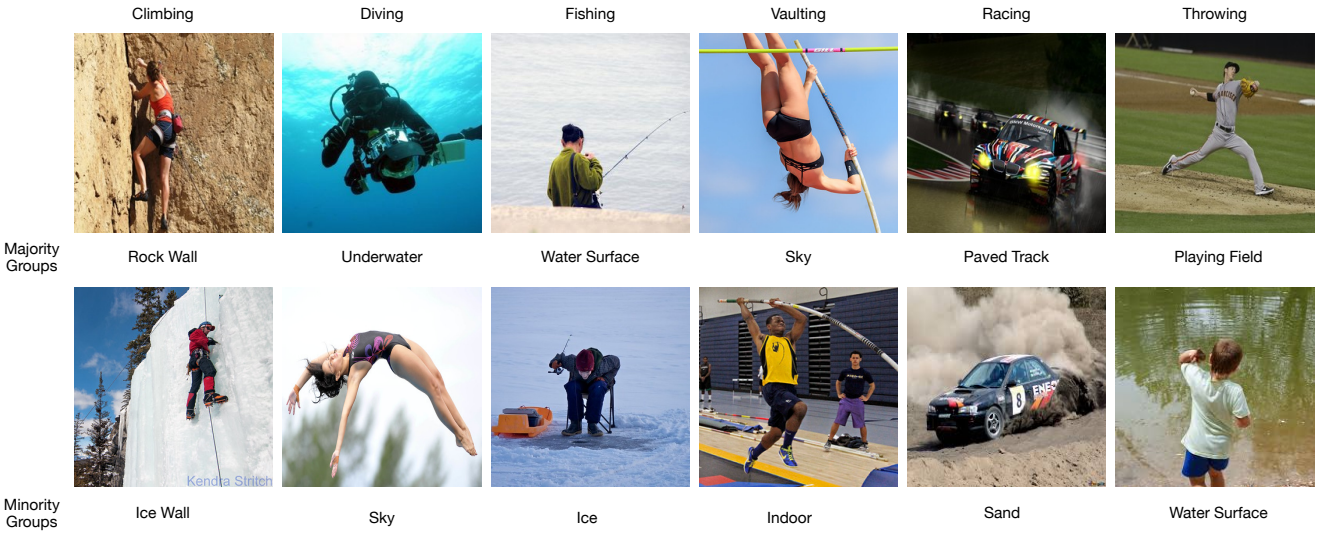

Figure 7. Image samples of BAR dataset.