Sufficient Invariant Learning for Distribution Shift

Taero Kim[1], Subeen Park[1], Sungjun Lim[1], Yonghan Jung[2], Krikamol Muandet[3], Kyungwoo Song[1]
[1]Yonsei University, [2]Purdue University, [3]CISPA Helmholtz Center for Information Security
taero.kim@yonsei.ac.kr, sallyna602@yonsei.ac.kr, lsj9862@yonsei.ac.kr,
jung222@purdue.edu, muandet@cispa.de, kyungwoo.song@yonsei.ac.kr

Abstract

Learning robust models under distribution shifts between training and test datasets is a fundamental challenge in machine learning. While learning invariant features across environments is a popular approach, it often assumes that these features are fully observed in both training and test sets—a condition frequently violated in practice. When models rely on invariant features absent in the test set, their robustness in new environments can deteriorate. To tackle this problem, we introduce a novel learning principle called the Sufficient Invariant Learning (SIL) framework, which focuses on learning a sufficient subset of invariant features rather than relying on a single feature. After demonstrating the limitation of existing invariant learning methods, we propose a new algorithm, Adaptive Sharpness-aware Group Distributionally Robust Optimization (ASGDRO), to learn diverse invariant features by seeking common flat minima across the environments. We theoretically demonstrate that finding a common flat minima enables robust predictions based on diverse invariant features. Empirical evaluations on multiple datasets, including our new benchmark, confirm ASGDRO's robustness against distribution shifts, highlighting the limitations of existing methods. Code: https://github.com/MLAI-Yonsei/SIL-ASGDRO.

1. Introduction

Machine learning models typically assume that training and test data are drawn from the same distribution. However, in real-world scenarios, this assumption is often violated whenever the training and test distribution differ, known as distribution shifts. In these cases, model performance tends to degrade, highlighting the need to develop models that are robust to distribution shifts for reliable outcomes.

To train models robust to distribution shift, invariant learning focuses on identifying latent features that remain constant across environments, referred to as invariant features. These features enable consistent predictions across environments by discouraging models from relying on spurious features [3] – features that are not preserved across changes in environments or groups[1]. For example, in domain generalization tasks [14, 20], the goal is to learn invariant features that consistently predict labels across multiple environments. Assuming that the learned invariant features persist in all unseen environments, they guarantee the model's generalization performance on new environments [23, 26]. Similarly, learning models robust to subpopulation shifts is essential in cases of severe imbalances between groups. In this scenario, invariant features play a crucial role in addressing the challenges faced by underrepresented groups, which are disproportionately impacted by strong spurious correlations [17, 35, 41].

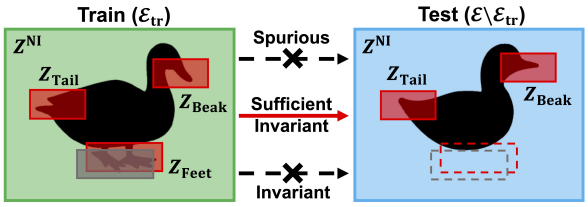

Figure 1. Left visualizes the images that contain a spurious feature, Z^{NI}, and multiple invariant features, Z_{Tail}, Z_{Beak}, and Z_{Feet} in training environment $\mathcal{E}_{tr}$. If the model focuses on the Z^{NI} (green background), then it fails to predict correctly in the test environment $\mathcal{E}\backslash\mathcal{E}_{tr}$ (Right). Even if the model captures the invariant features in $\mathcal{E}_{tr}$, e.g., Z_{Feet}, it still fails to predict correctly when the invariant features are not present (Gray). However, it is possible to predict correctly if we learn diverse invariant features sufficiently, Z_{Feet}, Z_{Tail}, and Z_{Beak}. With SIL (Red), the model predicts the label using remaining invariant features, Z_{Tail} and Z_{Beak} even though Z_{Feet} is not present in the test environment $\mathcal{E}\backslash\mathcal{E}_{tr}$.

However, learning all possible invariant features is challenging in practice because most existing invariant learning approaches focus on eliminating spurious correlations, which can be achieved by leveraging only a subset of the invariant features present in the training environments. Moreover, invariant features identified by the model may not be observable in unseen environments [15, 38]. This under-

[1]In this paper, the terms *environment* and *domain* are used interchangeably. A *group* refers to a subpopulation corresponding to a particular label within a specific environment.

scores the importance of learning a *sufficient* number of invariant features, rather than relying on a single invariant feature. To address this, we introduce a novel approach called *Sufficient Invariant Learning* (SIL), which focuses on learning a sufficient set of invariant features for improved generalization. For example, consider the scenario depicted in Fig. 1. Training environments for an image of a bird may include multiple invariant features, such as Z_{Tail}, Z_{Beak} and Z_{Feet}. If a model relies on a single invariant feature, say Z_{Feet}, it may fail to classify an image of the bird if the feature is unobservable (e.g., the bird's feet are hidden underwater). In contrast, if the model uses a sufficiently diverse set of invariant features (e.g., all of Z_{Tail}, Z_{Beak} and Z_{Feet}), it can still classify the image correctly as long as one or more of the other invariant features are present. This highlights the robustness and generalization benefits of learning a sufficient number of invariant features.

In this study, we develop the SIL framework and demonstrate that leveraging sufficiently diverse invariant features through SIL enhances model robustness. As a method for SIL, we propose *Adaptive Sharpness-aware Group Distributionally Robust Optimization* (ASGDRO). We show that ASGDRO attains SIL by effectively learning diverse invariant features while successfully eliminating spurious correlations. Furthermore, we show that the ability of ASGDRO to perform SIL is due to its convergence to a common flat minima [13] across diverse environments. Through empirical evaluations on a toy example and our newly introduced SIL benchmark dataset, we show that existing invariant learning algorithms fall short in capturing diverse invariant features, whereas ASGDRO successfully achieves SIL. By learning a wide range of invariant features sufficiently, ASGDRO exhibits robust generalization performance under various distribution shift scenarios, as evidenced by extensive experiments involving subpopulation and domain shifts.

2. Related Works

2.1. Invariant Learning for Distribution Shift

The standard approach to modern deep learning is Empirical Risk Minimization (ERM) [39], which minimizes the average training loss. However, ERM may not guarantee robustness in distribution shifts. To improve the generalization performance in distribution shift, Group Distributionally Robust Optimization (GDRO) minimizes the worst group loss for each iteration to alleviate spurious correlations [35]. Meanwhile, various studies utilize loss gradient for invariant learning. For example, Arjovsky et al. [3] minimizes the gradient norm of the fixed classifier across environments. Other research matches the loss gradient for each environment to find invariant features [31, 36]. Furthermore, balancing the representation using selective sampling with mix-up samples [41] or re-training the classifier on a small balanced set [19] show the effectiveness of learning a robust model. Some studies enhance generalization by combining invariant learning algorithms with feature extractors with rich representations [8, 43, 44] or resolving the conflict between ERM and invariant learning objectives [7].

Under the assumption that invariant features in the training environment also exist in the test environment, invariant learning theoretically guarantees an optimal predictor [34]. However, we argue that existing invariant learning algorithms do not learn sufficiently diverse invariant features, and they still suffer significant performance drops in test environments where some invariant features are unobserved [15, 38]. Lin et al. [24] consider settings with multiple features; however, they solely address scenarios where only spurious features are multiple in nature. To remedy this problem, we introduce the novel framework, SIL, and guarantee the generalization ability for diverse invariant features. Through experiments on the newly proposed benchmark in this paper, as well as on existing benchmarks for distribution shifts [14, 20], we demonstrate that our novel algorithm designed for SIL leads to more robust predictions.

2.2. Flatness and Generalization

Various studies argue that finding flat minima improves generalization performance [18, 27]. As a result, many algorithms emerge to find flat minima. Sharpness-aware Minimization (SAM) [13] finds flat minima by minimizing the maximum training loss of neighborhoods for the current parameter within ρ radius ball on the parameter space. Moreover, Adaptive SAM (ASAM) introduces the normalization operator to get a better correlation between flatness and the model's generalization ability by avoiding the scale symmetries between the layers [22]. Stochastic Weight Averaging (SWA) also reaches the flat minima by averaging the weight [17]. Under the IID setting, these approaches [13, 17, 22] successfully decrease the generalization gap.

Cha et al. [5] shows that optimizing the model towards flatter minima through weight averaging improves domain generalization ability. However, it is still necessary to verify whether the models operate robustly through weight averaging when strong spurious correlations exist. Indeed, some studies demonstrate that weight averaging may still not be robust in certain subpopulation shift tasks [32]. Zhang et al. [46] also shows that flat minima make the models more robust to the noise. However, our study focuses on the effectiveness of flatness in more extreme distribution shift settings, such as subpopulation shift and domain generalization. Springer et al. [37] presents that when easy-to-learn and hard-to-learn features coexist, models trained by SAM learn balanced representations. This aligns with our observations, and we aim to achieve SIL by removing spurious correlations and learning sufficiently diverse invariant features by introducing the constraints related to flatness.

3. Methodology

3.1. Problem Setting

Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, and Θ denote the input, label, feature, and parameter spaces, respectively. Consider a set of environments $\mathcal{E}$, where each environment $e \in \mathcal{E}$ is associated with a dataset $\mathcal{D}^e = \{(X_i^e, Y_i^e)\}_{i=1}^{n_e}$, with $X_i^e \in \mathcal{X}$, $Y_i^e \in \mathcal{Y}$, and n_e indicating the number of data points in e. We assume a feature set $Z = (Z^{\mathrm{I}}, Z^{\mathrm{NI}}) \subset \mathcal{Z}$, where Z^{I} denotes invariant features that satisfy the following invariance condition, and Z^{NI} denotes spurious features whose correlation with Y^e varies across environments e [3, 10, 21].

Definition 1 (Invariance Condition)**.** *Z^{I} is a set of invariant features satisfying*

$$\mathbb{E}[Y^e|Z^{\mathrm{I}}] = \mathbb{E}[Y^{e'}|Z^{\mathrm{I}}] \quad \textit{for all } e, e' \in \mathcal{E}_{\mathrm{tr}},$$

$\mathcal{E}_{\mathrm{tr}} \subset \mathcal{E}$ denotes the set of training environments.

We denote an invariant feature Z_i^{I} as the singleton set containing ith element of Z^{I}, where $i \in \{1, \ldots, p\}$ and p is the number of invariant features. In Fig. 1, the invariant features are $Z^{\mathrm{I}} = \{Z_{\text{Beak}}, Z_{\text{Tail}}, Z_{\text{Feet}}\}$ with $p = 3$, the spurious feature is $Z^{\mathrm{NI}} = \{Z_{\text{Background}}\}$ and one example of an invariant feature is $Z_1^{\mathrm{I}} = \{Z_{\text{Feet}}\}$, corresponding to the feet.

Suppose a model $f = h \circ g$ parametrized by $\theta = (\theta_g, \theta_h) \in \Theta$, where $g : \mathcal{X} \to \mathcal{Z}$ is an encoder with parameters θ_g and $h : \mathcal{Z} \to \mathcal{Y}$ is a classifier with parameters θ_h. Let $\mathcal{R}^e(\theta) = \mathbb{E}[\ell(f(X^e;\theta), Y^e)]$ denote the risk of a model f in environment e, where ℓ denotes a loss function. Invariant learning seeks to minimize the maximum risk across environments,

$$\min_{\theta} \max_{e \in \mathcal{E}} \mathcal{R}^e(\theta), \tag{1}$$

and to train models that have robust performance and generalization ability for unseen environments by learning invariant features Z^{I} [3, 10, 21, 35]. In particular, given Z^{I}, Rojas-Carulla et al. [34] demonstrate that learning optimal classifier θ_h^*, which is based on all invariant features in Z^{I}, leads to robust model predictions, i.e.,

$$\theta_h^* \in \min_{\theta_h} \max_{e \in \mathcal{E}} \mathcal{R}^e(\theta_h), \tag{2}$$

where $\mathcal{R}^e(\theta_h) = [\ell(h(Z^{\mathrm{I}};\theta_h), Y^e)]$, assuming that the invariance condition holds for all $e \in \mathcal{E}$.

3.2. Sufficient Invariant Learning

While models trained via invariant learning have shown effectiveness under various distribution shifts, this does not imply that the optimal classifier satisfying Eq. (2) is unique. For $\mathcal{E}_{\mathrm{tr}}$, Definition 1 holds for any subset $\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}}$ and any classifier relying on $\hat{Z}^{\mathrm{I}}$ can be optimal. In Fig. 1, the classifier may utilize only Z_{Feet}, or it may employ all Z_i^{I} simultaneously in $\mathcal{E}_{\mathrm{tr}}$, to distinguish between waterbirds from landbirds. Therefore, the optimal encoder minimizing Eq. (1) is also not unique, as it depends on the non-unique optimal classifier [3]. To distinguish predictive mechanisms using different $\hat{Z}^{\mathrm{I}}$, we define the invariant mechanism.

Definition 2 (Invariant Mechanism)**.** *For an encoder $g_{\theta_g^{\mathrm{I}}}$ parameterized by θ_g^{I} and a classifier $h_{\theta_h^{\mathrm{I}}}$ parameterized by θ_h^{I}, the invariant mechanism $\theta^{\mathrm{I}} = (\theta_g^{\mathrm{I}}, \theta_h^{\mathrm{I}}) \in \Theta$ is a tuple for a subset $\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}}$ satisfying the followings:*

$$\textit{Condition 1: } h_{\theta_h^{\mathrm{I}}} : \hat{Z}^{\mathrm{I}} \mapsto Y^e, \quad \forall e \in \mathcal{E}_{\mathrm{tr}}.$$
$$\textit{Condition 2: } \theta^{\mathrm{I}} \in \operatorname*{argmin}_{\theta} \max_{e \in \mathcal{E}_{\mathrm{tr}}} \mathcal{R}^e(\theta).$$

Specifically, we denote the invariant mechanism that utilizes only Z_i^{I} as θ_i^{I}, for $i = \{1, \ldots p\}$. Invariant mechanisms that rely solely on a specific invariant feature θ_i^{I} may struggle to make robust predictions when the part of the input corresponding to that feature is corrupted by noise, missing due to cropping, or occluded by environmental factors. This non-uniqueness suggests that training encoders via classifier invariance [1, 3] or enhancing them to capture richer information [8, 44] can benefit from additional regularization to leverage other invariant features. This observation also implies that robust optimization methods designed to minimize Eq. (1) over $\mathcal{E}_{\mathrm{tr}}$ [12, 30, 35] have an avenue for achieving enhanced generalization performance.

We argue that training more robust models requires ensuring generalization across sufficiently diverse sets of invariant features. To this end, we introduce a novel invariant learning framework, termed Sufficient Invariant Learning (SIL), which encourages learning diverse invariant features:

Definition 3 (Sufficient Invariant Learning)**.** *Sufficient Invariant Learning refers to identify θ^{SI} such that*

$$\theta^{\mathrm{SI}} \in \operatorname*{argmin}_{\theta} \max_{e \in \mathcal{E}} \mathcal{R}^e(\theta),$$
$$s.t. \quad \theta_h^{\mathrm{SI}} \in \operatorname*{argmin}_{\theta_h} \max_{e \in \mathcal{E}} \max_{\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}}} \mathbb{E}[\ell(h_{\theta_h}(\hat{Z}^{\mathrm{I}}), Y^e)].$$

SIL aims to train a classifier that performs robustly not only across all environments but also with respect to any subset $\hat{Z}^{\mathrm{I}}$. It encourages the model to leverage sufficiently diverse invariant features, assuming that representations of these features have already been learned from the target task [19]. The main challenge in achieving SIL lies in the cost of obtaining individually intervened data for each $\hat{Z}^{\mathrm{I}}$. To achieve this, we propose ASGDRO, a novel method inspired by the geometry of the loss surface, which promotes SIL by identifying common flat minima.

3.3. ASGDRO: Adaptive Sharpness-aware Group Distributionally Robust Optimization

In the literature on model merging and multi-task learning [2, 16, 33, 40], it is often assumed that a robust model across all tasks lies within the linear interpolation of models that perform well on each individual task. Inspired by this observation, we consider θ_i^{I} as a model that performs well on a single task, and we hypothesize that θ^{SI} exists within the linear interpolation of these mechanisms. Without loss of generality, subsets that are not singletons can be equivalently represented as an interpolation of singleton invariant features Z_i^{I}. Hence, for the remainder of this work, we restrict our consideration to Z_i^{I} and Z^{I} (Appendix A.2.). The key difference from previous studies is that we evaluate each task solely on the same dataset. Therefore, as discussed in Sec. 3.2, different invariant mechanisms are expected to have similar risks,

$$\mathcal{R}^e(\theta^{\mathrm{SI}}) - \mathcal{R}^e(\theta_i^{\mathrm{I}}) \approx 0 \quad \text{for all } e \in \mathcal{E}_{\mathrm{tr}}.$$

A challenge for SIL is that we do not have access to information about θ_i^{I}. However, based on the observation in Neyshabur et al. [28] that different models trained from the same pre-trained model lie in the same loss basin, we assume that models located on the linear path between θ^{SI} and θ_i^{I} also exhibit similar risk. Therefore, θ^{SI} should guarantee low risks within a ball of radius at least $\max_i ||\theta_i^{\mathrm{I}} - \theta^{\mathrm{SI}}||$, denoted as ρ, in Euclidean space. Introducing a perturbation $\epsilon_e := \theta_i^{\mathrm{I}} - \theta^{\mathrm{SI}}$, we obtain the following condition for the risk of θ_i^{I}:

$$\max_{i\in\{1,\dots,p\}} \mathcal{R}^e(\theta_i^{\mathrm{I}}) = \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta^{\mathrm{SI}} + \epsilon_e).$$

From our motivation, ρ is a hyper-parameter adjusting the model class of θ_i^{I} deviated from θ^{SI}. Moreover, according to Definition 1, all θ_i^{I} should exhibit robust performance across environments $e \in \mathcal{E}_{\mathrm{tr}}$. Finally, we propose a novel objective function named Adaptive Sharpness-aware Group Distributionally Robust Optimization (ASGDRO), which is formulated as follows:

$$\max_{e\in\mathcal{E}_{tr}} \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta + \epsilon_e). \tag{3}$$

In the following sections, we theoretically show that ASGDRO not only learns invariant features but also balances the learning of invariant mechanisms, thereby achieving SIL. Also, we demonstrate that ASGDRO finds the common flat minima across environments, leading to SIL.

3.4. SIL and Common Flat Minima

We demonstrate that ASGDRO trains the model to achieve SIL by showing that ASGDRO balances the use of diverse invariant mechanisms.

Algorithm 1 ASGDRO

Input: Training dataset $D_{\mathrm{tr}}^e = \{(X^e, Y^e)\}$ for $e \in \mathcal{E}_{\mathrm{tr}}$, Radius $\rho > 0$, Learning rate $\eta > 0$, Robust step size $\gamma > 0$, The number of environments $|\mathcal{E}_{\mathrm{tr}}|$, Normalization Matrix T_θ.

Initialization: θ_0; $\lambda_e^{(0)} = 1/|\mathcal{E}_{\mathrm{tr}}|$;
for $t = 1, 2, 3, \dots$ **do**
  for $e = 1, \dots, |\mathcal{E}_{\mathrm{tr}}|$ **do**
    Compute training loss $\mathcal{R}^e(\theta_t)$;
    Compute $\epsilon_e^* = \rho \frac{T_\theta^2 \nabla \mathcal{R}^e(\theta_t)}{\|T_\theta \nabla \mathcal{R}^e(\theta_t)\|}$;
    Gradient ascent: $\theta_t^* = \theta_t + \epsilon_e^*$;
    Find loss for each environment $\mathcal{R}^e(\theta_t^*)$;
    Compute $\tilde{\lambda}_e^{(t)} = \lambda_e^{(t-1)} \exp(\gamma \mathcal{R}^e(\theta_t^*))$;
    Return to θ_t;
  end for
  Update $\lambda_e^{(t)} = \tilde{\lambda}_e^{(t)} / \sum_e \tilde{\lambda}_e^{(t)}$;
  Compute $\mathcal{R}_{\mathrm{ASGDRO}}(\theta_t) = \sum_e \lambda_e^{(t)} \mathcal{R}^e(\theta_t^*)$;
  Compute $\nabla \mathcal{R}_{\mathrm{ASGDRO}}(\theta_t) = \sum_e \lambda_e^{(t)} \nabla \mathcal{R}^e(\theta_t^*)$;
  Return to θ_t;
  Update the parameters: $\theta_{t+1} = \theta_t - \eta \nabla \mathcal{R}_{\mathrm{ASGDRO}}(\theta_t)$;
end for

Theorem 1. *Let $\theta_\lambda^{\mathrm{I}}$ be a convex combination of θ_i^{I}, where λ is a p-dimensional vector. Consider mean-squared error as the loss function. Assume a linear model with $Z \in \mathbb{R}^p$, where the p features are orthogonal, and suppose $Z = Z^{\mathrm{I}} = (1, \dots, 1)$. Then,*

$$\begin{aligned}\lambda^* &= \operatorname*{argmin}_\lambda \max_{e\in\mathcal{E}_{tr}} \max_{||\epsilon||\leq\rho} \mathcal{R}^e(\theta_\lambda^{\mathrm{I}} + \epsilon) \\ &\approx \operatorname*{argmin}_\lambda \max_{e\in\mathcal{E}_{tr}} \left[\mathcal{R}^e(\theta_\lambda^{\mathrm{I}}) + \rho||\lambda|| \cdot ||\nabla \mathcal{R}^e(\theta_\lambda^{\mathrm{I}})||\right] \\ &= \operatorname*{argmin}_\lambda ||\lambda|| = (\frac{1}{p}, \dots, \frac{1}{p})\end{aligned} \tag{4}$$

where $||\cdot||$ denotes L_2 norm.

Refer to Appendix A.4. for the proof. Theorem 1 states that ASGDRO ensures that even when invariant features contribute equally to the output, the model does not favor a simple solution focusing on a single invariant feature. Instead, it learns a diverse range of invariant mechanisms. As shown in Eq. (4), this regularization effect arises through the gradient norm $||\nabla \mathcal{R}^e(\theta)||$.

Proposition 1. *By the Taylor expansion,*

$$\max_{e\in\mathcal{E}} \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta + \epsilon_e) \approx \max_{e\in\mathcal{E}} [\mathcal{R}^e(\theta) + \rho||\nabla \mathcal{R}^e(\theta)||].$$

ASGDRO leads to a regularization of the gradient norm, $\mathcal{R}^e$, $||\nabla \mathcal{R}^e(\theta)||$, across environments, which drives the model to converge to common flat minima.

Refer to Appendix A.3. for proof. As demonstrated in [47], small $||\nabla \mathcal{R}^e(\theta)||$ indicates flat minima. We also demonstrate this property empirically in Fig. 5 and Appendix

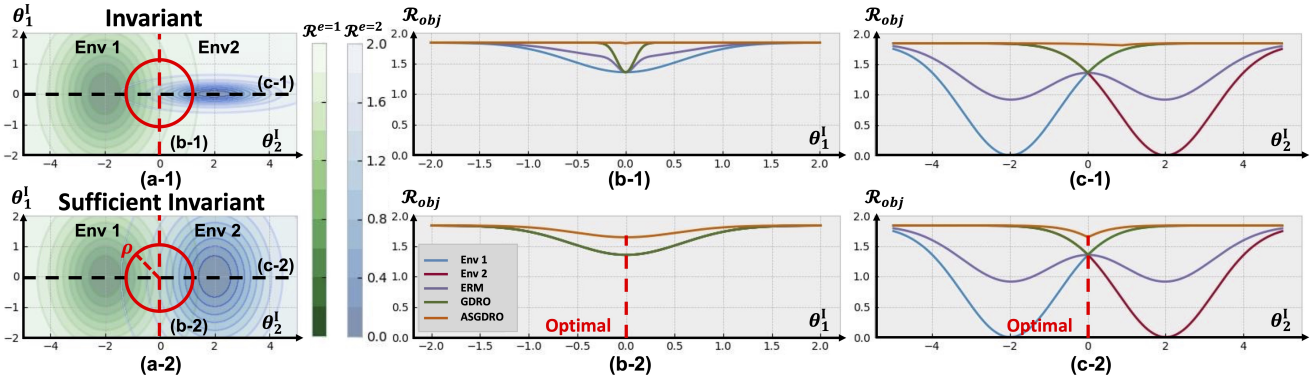

Figure 2. **Sufficient Invariant Learning and Common Flat Minima** In (a-1) and (a-2), two axes, θ_1^{I} and θ_2^{I}, represent the invariant directions of parameters corresponding to each invariant mechanism respectively. The red circle indicates the area bound by ρ for measuring flatness in ASGDRO. (b-1) and (b-2) show that when Env 2 has sharp minima in the direction of θ_1^{I}, GDRO still converges, but ASGDRO does not have any optimal point due to the sharpness of θ_1^{I}. However, in (c-1) and (c-2) when both invariant directions of Env 2 as well as Env 1 are flat, ASGDRO has an optimal point and prefers to converge. That is, ASGDRO learns diverse invariant features sufficiently.

A.11. Finally, we argue that finding common flat minima encourages the model to learn sufficiently diverse invariant mechanisms. Moreover, this aligns with existing studies in IID settings, which suggest that flatter minima improve the generalization performance of models [13, 18, 22]. Additionally, we demonstrate in Appendix A.5. that ASGDRO successfully eliminates the spurious feature Z^e while effectively learning the invariant feature.

3.5. Implementation of ASGDRO

From Foret et al. [13], maximum value of inner term in Eq. (3) is approximated when $\epsilon_e = \rho \frac{\nabla \mathcal{R}^e(\theta)}{\|\nabla \mathcal{R}^e(\theta)\|}$. However, Kwon et al. [22] show that by introducing the normalization matrix T_θ, which removes the scale symmetry present on the loss surface, the correlation between flatness and generalization performance is strengthened. ASGDRO also adopts the same T_θ, and modified objective function is as follows:

$$\mathcal{R}_{\mathrm{ASGDRO}}(\theta) = \max_{e \in \mathcal{E}_{\mathrm{tr}}} \mathcal{R}^e(\theta + \epsilon_e^*),$$

where $\epsilon_e^* = \rho \frac{T_\theta^2 \nabla \mathcal{R}^e(\theta)}{\|T_\theta \nabla \mathcal{R}^e(\theta)\|}$ is an adversarial perturbation for each environment e, $T_\theta = \mathrm{diag}(\mathrm{concat}(\|\mathbf{k_1}\|\mathbf{1}_{\mathbf{n(k_1)}}, \ldots, \|\mathbf{k_m}\|\mathbf{1}_{\mathbf{n(k_m)}}, |\omega_1|, \ldots, |\omega_q|))$, where $\mathbf{k_m}$ denotes a convolution kernel, $\omega_\mathbf{q}$ represents other parameters and $\mathbf{n}(\cdot)$ indicates the number of parameters.

To address the instability in training that arises from the optimization approach of selecting only the worst environment at each step, we adopt an alternative gradient-based optimization algorithm inspired by GDRO [35]. We modify ASGDRO into the form of linear interpolation across environments and update their coefficients:

$$\max_{e \in \mathcal{E}_{\mathrm{tr}}} \mathcal{R}^e(\theta + \epsilon_e^*) = \max_{\sum_e \lambda_e = 1, \lambda_e \geq 0} \sum_{e \in \mathcal{E}_{\mathrm{tr}}} \lambda_e \mathcal{R}^e(\theta + \epsilon_e^*),$$

where λ_e is the weight imposed on adversarial perturbed loss for each environment. Finally, we update our model parameter from the current parameter θ_t as follows:

$$\theta_t - \eta \nabla \mathcal{R}_{\mathrm{ASGDRO}}(\theta) = \theta_t - \eta \sum_{e \in \mathcal{E}_{\mathrm{tr}}} \lambda_e^{(t)} \nabla \mathcal{R}^e(\theta_t + \epsilon_e^*),$$

where η denotes the learning rate and $\lambda_e^{(t)}$ denotes the weight imposed on each loss of environment at time step t. Refer to Algorithm 1 for the details. In practice, for computational efficiency, in all experiments except for the toy example, instead of calculating $\epsilon_e^* = \rho \frac{T_\theta^2 \nabla \mathcal{R}^e(\theta)}{|T_\theta \nabla \mathcal{R}^e(\theta)|}$ for each environment, we use a common adversarial perturbation utilizing the empirical risk $\mathcal{R}_S(\theta) = \frac{1}{|D^e||\mathcal{E}_{\mathrm{tr}}|} \sum_{e \in \mathcal{E}_{\mathrm{tr}}} \sum_{n_e} \ell(f(X^e; \theta), Y^e)$, i.e. $\epsilon^* = \rho \frac{T_\theta^2 \nabla \mathcal{R}_S(\theta)}{|T_\theta \nabla \mathcal{R}_S(\theta)|}$. As a result, the gradient ascending through ϵ^* is performed only once regardless of the number of environments, and the loss for each environment is evaluated using the same perturbed parameters, $\theta + \epsilon^*$.

4. Experiments

4.1. Toy Exmaple

We demonstrate through a toy example that the representative invariant learning algorithm GDRO [35] fails to learn diverse invariant mechanisms, whereas ASGDRO successfully achieves SIL by encouraging the model to converge to the common flat minima (Fig. 2). First, we assume that we know two different directions corresponding to the different invariant mechanism θ_1^{I} and θ_2^{I}, which learns different invariant features, Z_1^{I} and Z_2^{I}, respectively. We define the loss surface of each environment e following a Gaussian function with respect to θ_1^{I} and θ_2^{I}:

$$G(\theta) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)\right),$$

$$\text{where} \quad \theta = \begin{bmatrix} \theta_1^{\mathrm{I}} \\ \theta_2^{\mathrm{I}} \end{bmatrix}, \mu^{(e)} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma^{(e)} = \begin{bmatrix} \sigma_{11} \sigma_{12} \\ \sigma_{21} \sigma_{22} \end{bmatrix}.$$

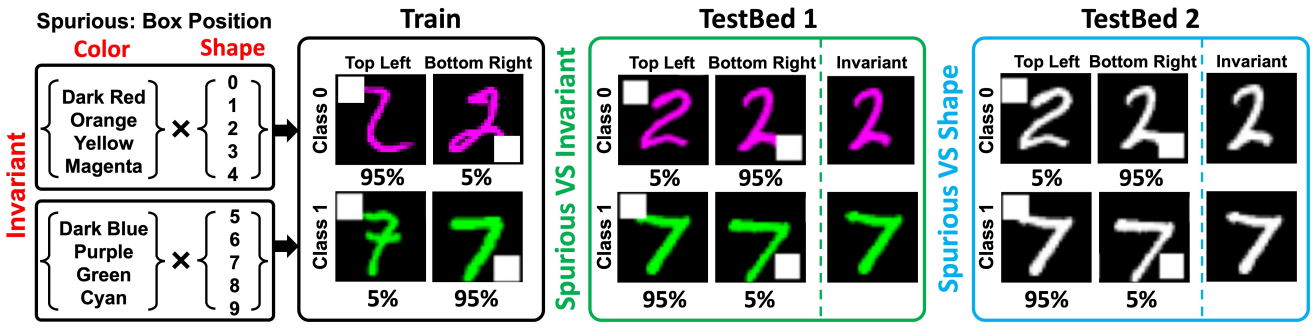

Figure 3. **Overview of H-CMNIST.** There are three features, color and shape (invariant features, $Z^{\mathrm{I}} = \{Z_{\mathrm{color}}, Z_{\mathrm{shape}}\}$) and box position (spurious feature, $Z^{\mathrm{NI}} = \{Z_{\mathrm{BP}}\}$). The ratio of Z_{BP} is flipped between the train and test set. The test set consists of two testbeds, one for evaluating whether learning invariant features and the other for evaluating whether learning sufficiently diverse invariant features.

	TestBed 1		TestBed 2	
	Spu & Inv	Inv	Spu & Shape	Shape
ERM	97.11 ± 3.44	98.75 ± 1.19	34.64 ± 9.90	57.41 ± 2.58
ASAM	98.57 ± 1.21	98.12 ± 1.74	34.78 ± 8.41	57.07 ± 1.91
GDRO	$\mathbf{99.95 \pm 0.07}$	$\mathbf{99.92 \pm 0.08}$	57.53 ± 2.11	61.44 ± 1.03
ASGDRO	99.88 ± 0.11	99.83 ± 0.12	$\mathbf{66.62 \pm 5.61}$	$\mathbf{69.17 \pm 6.19}$

Table 1. **H-CMNIST Results.** TestBed 1 evaluates whether the model learns easy invariant feature Z_{color}, and TestBed 2 evaluates the ability to learn additional invariant feature Z_{shape}.

To make losses greater than 0, we subtracted $G(\theta)$ from its maximum value. As a result, we define the loss surface corresponding to the two environments, each with a minimum value of 0, as follows:

$$\mathcal{R}^{e=1}(\theta) = \max_{\theta} \left[G(\theta; \mu^{(1)}, \Sigma^{(1)}) \right] - G(\theta; \mu^{(1)}, \Sigma^{(1)})$$
$$\mathcal{R}^{e=2}(\theta) = \max_{\theta} \left[G(\theta; \mu^{(2)}, \Sigma^{(2)}) \right] - G(\theta; \mu^{(2)}, \Sigma^{(2)})$$

Now we create sharp or flat minima in a specific direction by adjusting the covariance matrix $\Sigma^{(e)}$. In this example, we consider a fixed situation where both $e = 1$ and $e = 2$ have flat minima with respect to θ^{I}_2. When $\mathcal{R}^{e=1}(\theta)$ always has flat minima in the direction of θ^{I}_1, we aim to observe how the loss $\mathcal{R}_{obj}$ corresponding to each objective function changes depending on whether θ^{I}_2 has sharp or flat minima (a-1 and a-2 in Fig. 2). The parameters that we use to generate the toy examples are as follows:

$$\text{Env 1 } (e=1): \ \mu = \begin{bmatrix} -2.0 \\ 0.0 \end{bmatrix}, \ \text{Env 2 } (e=2): \ \mu = \begin{bmatrix} 2.0 \\ 0.0 \end{bmatrix}$$
$$\text{Flat}: \ \Sigma = \begin{bmatrix} 1.5 & 0.0 \\ 0.0 & 2.0 \end{bmatrix}, \ \text{Sharp}: \ \Sigma = \begin{bmatrix} 1.5 & 0.0 \\ 0.0 & 0.05 \end{bmatrix}$$

We evaluate each algorithm through the loss surface in each direction (second and third columns of Fig. 2). When Env 2 exhibits sharpness for θ^{I}_1 (first row of Fig. 2), it indicates that learning the invariant feature corresponding to θ^{I}_1 may result in a large generalization gap [18]. However, GDRO does not incorporate regularization on flatness and only considers the loss at the current parameter, allowing convergence to a sharp solution. From Theorem 1, it implies the large gradient norm, and this situation does not constitute successful SIL. In contrast, ASGDRO, which takes into account the loss in neighboring parameters, avoids sharp regions for θ^{I}_1 (b-1 and c-1 in Fig. 2).

When Env 2 is flat for θ^{I}_1 (second row in Fig. 2), we say that the model performs SIL if it converges into the common flat minima between Env 1 and Env 2. However, GDRO has the same loss at the optimal point in this situation as in the previous case, indicating that GDRO does not specifically regularize the model to perform SIL. On the other hand, ASGDRO, by accounting for common flat minima, identifies an optimal parameter that promotes learning of diverse invariant mechanisms (b-2 and c-2 in Fig. 2). As a result, by considering flatness, the model performs SIL and is expected to make robust predictions in unseen environments by leveraging multiple invariant features.

4.2. Heterogenous ColoredMNIST

By finding the common flat minima, ASGDRO learns diverse invariant features. To demonstrate this, we propose Heterogeneous ColoredMNIST (H-CMNIST), a new dataset designed to evaluate whether the model learns diverse invariant mechanisms sufficiently (Fig. 3). H-CMNIST evaluate whether the remaining invariant feature is additionally learned by the algorithm, assuming that the model has already learned one invariant feature.

H-CMNIST includes two invariant features, the color $Z^{\mathrm{I}}_1 = \{Z_{\mathrm{color}}\}$ and shape of digits $Z^{\mathrm{I}}_2 = \{Z_{\mathrm{shape}}\}$, and one spurious feature, the position of the box (BP) $Z^{\mathrm{NI}} = \{Z_{\mathrm{BP}}\}$. That is, each class has its own colors and shapes. Using BP, we construct two environments, Top Left (Env 0) and Bottom Right (Env 1). We design a scenario where spurious correlations occur [11, 35]. Specifically, in the training set, 95% of Left Top BP belongs to class 0, and only 5% belongs to class 1. In contrast, we collect 95% of Right Bottom BP in class 1, and assigned only 5% to class 0.In the test sets,

	CMNIST		Waterbirds		CelebA		CivilComments	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
ERM‡	27.8%	0.0%	97.0%	63.7%	94.9%	47.8%	92.2%	56.0%
ASAM	40.5%	34.1%	97.4%	72.4%	93.7%	46.5%	92.3%	58.9%
IRM‡	72.1%	70.3%	87.5%	75.6%	94.0%	77.8%	88.8%	66.3%
IB-IRM‡	72.2%	70.7%	88.5%	76.5%	93.6%	85.0%	89.1%	65.3%
V-REx‡	71.7%	70.2%	88.0%	73.6%	92.2%	86.7%	90.2%	64.9%
CORAL‡	71.8%	69.5%	90.3%	79.8%	93.8%	76.9%	88.7%	65.6%
GDRO‡	72.3%	68.6%	91.8%	90.6%	92.1%	87.2%	89.9%	70.0%
DomainMix‡	51.4%	48.0%	76.4%	53.0%	93.4%	65.6%	90.9%	63.6%
Fish‡	46.9%	35.6%	85.6%	64.0%	93.1%	61.2%	89.8%	71.1%
LISA‡	74.0%	73.3%	91.8%	89.2%	92.4%	89.3%	89.2%	**72.6%**
ASGDRO	74.8%	**74.2%**	92.3%	**91.4%**	92.1%	**91.0%**	90.2%	71.8%

Table 2. **Subpopulation Shift**. ‡ denotes the performance reported from [41]. Avg. denotes average accuracy, and Worst denotes worst group accuracy. Refer to Appendix A.7 for details.

the composition of BP is flipped. That is, Z_{BP} has a strong correlation with each class in the training set, but it does not hold in the test set. Refer to Appendix A.6 for details.

Tab. 1 shows the results of H-CMNIST. H-CMNIST assumes an easily learnable invariant feature Z_{color} to evaluate whether the model, having already learned one invariant feature, can learn additional invariant features Z_{shape}. Concretely, TestBed 1 serves as a preliminary step to verify that an easily learnable invariant feature is indeed present. In TestBed1, the performance of all algorithms is similar regardless of the presence of the spurious feature Z_{BP}, indicating that all have learned at least one invariant feature.

However, in Testbed 2, without Z_{color}, both ERM and ASAM show significant performance discrepancies depending on the presence of spurious feature Z_{BP}. Compared with the results of TestBed 1, ERM and ASAM only learn Z_{color} successfully, but they fail to capture the additional invariant feature, Z_{shape}. It indicates that even when a relatively easier invariant feature exists, the spurious feature influences the relatively more challenging invariant feature. Although GDRO exhibits robustness to spurious correlations compared to ERM and ASAM, it still fails to learn one of the invariant features, Z_{shape}. However, ASGDRO makes robust predictions against spurious features and more successful learning of shape features in TestBed2, compared to other baselines. It implies that SIL is necessary for the robust model and ASGDRO optimizes the model to learn sufficiently diverse invariant features $Z^{\text{I}} = \{Z_{\text{color}}, Z_{\text{shape}}\}$ considering the common flat minima across environments.

4.3. Experimental Results

In each result, boldface and underlined text denote the highest and second-highest accuracy for each dataset, respectively. Additional experiments, including efficiency or sensitivity analysis of ASGDRO can be found in the Appendix.

We conduct experiments for subpopulation shift, CMNIST [3], Waterbirds [35], CelebA [25], and CivilComments [4]. The goal of the subpopulation shift task is to obtain the better worst group performance by learning invariant features. Different from H-CMNIST, the spurious correlation acts as a stronger shortcut. As a result, the models cannot learn any invariant feature easily. Tab. 2 shows the results of subpopulation shift experiments. ASAM, which considers flatness, fails to eliminate spurious correlations and shows limited predictive accuracy on the worst group. On the other hand, ASGDRO shows the best and worst group performance for all data except CivilComments. For CivilComments data, ASGDRO also shows comparable performance with the best algorithms among the baselines. Compared to GDRO, the primary distinction of ASGDRO is its ability to find a common flat minima, which not only enhances robustness for the worst group but also reduces the gap between average accuracy and worst group accuracy. Therefore, Tab. 2 provides support for our claim that sufficiently learning diverse invariant mechanisms leads to robust generalization performance.

PT–FT	Camelyon17 Avg. (%)	CivilComments Worst (%)	FMoW Worst (%)	Amazon 10th per. (%)	RxRx1 Avg. (%)
×–ERM	70.3 ±6.4	56.0 ±3.6	32.3 ±1.3	53.8 ±0.8	29.9 ±0.4
×–GDRO	68.4 ±7.3	70.0 ±2.0	30.8 ±0.8	53.3 ±0.0	23.0 ±0.3
×–IRM	64.2 ±8.1	66.3 ±2.1	30.0 ±1.4	52.4 ±0.8	8.2 ±1.1
ERM–ERM	74.3 ±6.0	55.5 ±1.8	33.6 ±1.0	51.1 ±0.6	30.2 ±0.1
ERM–GDRO	76.1 ±6.5	69.5 ±0.2	33.0 ±0.5	52.0 ±0.0	30.0 ±0.1
ERM–IRM	75.7 ±7.4	68.8 ±1.0	33.5 ±1.1	52.0 ±0.0	30.1 ±0.1
Bonsai–ERM	74.0 ±5.3	63.3 ±3.5	31.9 ±0.5	48.6 ±0.6	24.2 ±0.4
Bonsai–GDRO	72.8 ±5.4	70.2 ±1.3	33.1 ±1.2	42.7 ±1.1	23.0 ±0.5
Bonsai–IRM	73.6 ±6.2	68.4 ±2.0	32.5 ±1.2	47.1 ±0.6	23.4 ±0.4
FeAT–ERM	77.8 ±2.5	68.1 ±2.3	33.1 ±0.8	52.9 ±0.6	30.7 ±0.4
FeAT–GDRO	80.4 ±3.3	71.3 ±0.5	33.6 ±1.7	52.6 ±0.6	30.0 ±0.1
FeAT–IRM	78.0 ±3.1	70.3 ±1.1	34.0 ±0.7	52.9 ±0.6	30.0 ±0.2
×–ASGDRO	**81.0** ±3.8	**71.8** ±0.4	**35.0** ±0.3	**54.5** ±0.5	**32.2** ±0.2

Table 3. **Wilds Benchmark.** Out-of-distribution generalization performances on wilds benchmark with rich representation. The performances of the baseline models are the reported results from [20] and [8]. × indicates the absence of a pre-training process on the target dataset. Refer to Appendix A.8 for error bars.

Method	PACS	VLCS	OH	TI	DN	Avg
ERM†	85.5	77.5	66.5	46.1	40.9	63.3
IRM†	83.5	78.6	64.3	47.6	33.9	61.6
GDRO†	84.4	76.7	66.0	43.2	33.3	60.7
I-Mixup†	84.6	77.4	68.1	47.9	39.2	63.4
MMD†	84.7	77.5	66.4	42.2	23.4	58.8
SagNet†	86.3	77.8	68.1	48.6	40.3	64.2
ARM†	85.1	77.6	64.8	45.5	35.5	61.7
VREx†	84.9	78.3	66.4	46.4	33.6	61.9
RSC†	85.2	77.1	65.5	46.6	38.9	62.7
GSAM [48]	85.9	79.1	**69.3**	47.0	44.6	65.1
RDM [29]	**87.2**	78.4	67.3	47.5	43.4	64.8
RS-SCM [9]	85.8	77.6	68.8	47.6	42.5	64.4
LFME [6]	85.0	78.4	69.1	48.3	42.1	64.6
ASGDRO	86.7	**80.0**	69.2	**48.8**	**44.9**	**65.9**
DPLCLIP	96.6	79.0	82.7	45.4	59.1	72.6
DPLCLIP+GDRO	95.9	79.7	83.6	46.0	59.1	72.9
DPLCLIP+ASGDRO	**96.8**	**80.7**	**83.7**	**48.9**	**59.8**	**74.0**

Table 4. **DomainBed.** The symbol † indicates reported performance in Gulrajani and Lopez-Paz [14]. Refer to Appendix A.9 for error bars and experimental details.

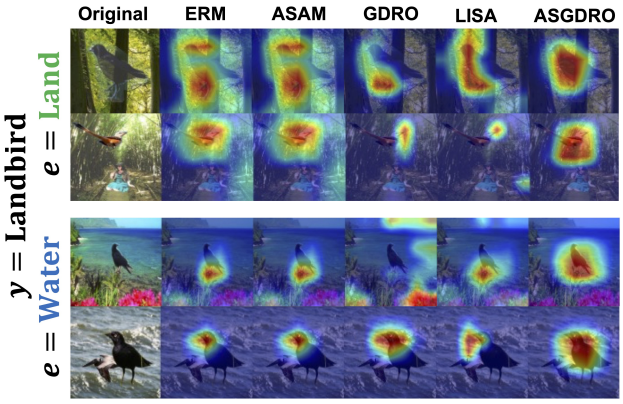

Figure 4. **Grad-CAM** ASGDRO learns diverse invariant features.

One approach to training a robust model is to enrich the representation learning of invariant features [8, 44] rather than training by ERM. This process consists of a pre-training (PT) stage dedicated to representation learning, followed by a fine-tuning (FT) stage utilizing existing invariant learning algorithms. In Tab. 3, we compare these algorithms with ASGDRO, evaluated on the Wilds benchmark dataset, which includes various types of distribution shifts collected from real-world scenarios. Notably, the superior performance of ASGDRO, even compared to invariant learning algorithms trained with rich representations during the FT stage, suggests that it is important not only to learn rich representations of invariant features but also to ensure that predictions are composed using diverse invariant features by learning sufficiently diverse invariant mechanisms.

We also conduct DomainBed benchmark [14], which is the most commonly used for evaluating domain generalization performance under a fair setting. ASGDRO is a model-agnostic method and is easily applied to various algorithms. Thus, we apply ASGDRO with DPLCLIP [45], which performs the prompt learning for domain generalization. Tab. 4 presents the performance of both the original ASGDRO and DPLCLIP when ASGDRO is applied. ASGDRO achieves the highest average performance compared to other algorithms. Additionally, for DPLCLIP, training with ASGDRO proves to be more effective across all datasets compared to training with standard ERM or GDRO.

4.4. Visual Interpretation by Grad-CAM

We conduct Grad-CAM analysis to verify whether the effect of learning SIL is being properly applied on the ground-truth label (Fig. 4). The minority group, land birds on a water background, is underrepresented by the spurious correlation as it has only a few samples. ERM and ASAM use several features to predict the majority group, land birds on a land background, but fail to remove spurious correlation. As a result, they also use the background feature. For the minority groups, however, only a small part of the invariant features is observed to be used for prediction. GDRO successfully removes spurious correlation regardless of the group but still uses only the part of invariant features for prediction. On the other hand, ASGDRO focuses on various invariant features for prediction regardless of the group; that is, it sufficiently uses diverse invariant features of land birds. Additionally, ASGDRO successfully excludes spurious features in their prediction. Appendix A.10. provides additional results on Grad-CAM.

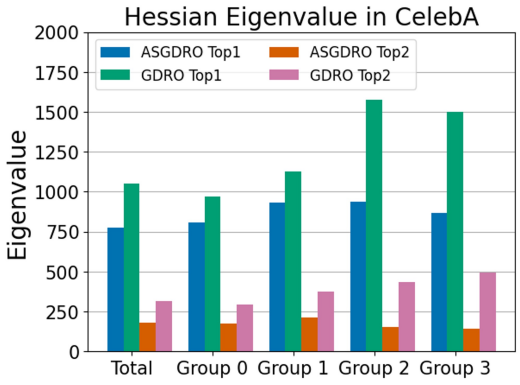

Figure 5. **Hessian Analysis on CelebA.** ASGDRO finds the common flat minima for all groups.

4.5. Hessian Analysis

In Fig. 5, we report the eigenvalues of the Hessian matrix to measure and compare the flatness of the model [42]. A lower eigenvalue indicates a flatter minima. Compared to GDRO, ASGDRO exhibits lower eigenvalues across all groups. Furthermore, GDRO shows particularly sharper minima in Group 2 and 3, which include minority groups. In contrast, ASGDRO maintains relatively uniform eigenvalues regardless of the group. This suggests that ASGDRO indeed finds a common flat minima, with the regularization for such minima enabling the model to make robust predictions by leveraging diverse invariant mechanisms. Refer to Appendix A.11. for additional experimental analysis.

5. Conclusion

This study highlights the significance of SIL, which promotes the learning of diverse invariant features. Unlike invariant learning, SIL enables models to leverage these diverse invariant mechanisms for prediction, ensuring robustness even in environments where some invariant features are unobserved. We also introduce ASGDRO, the first SIL algorithm designed to identify common flat minima across environments. Through both theoretical analysis and experimental validation, we demonstrate that ASGDRO effectively learns diverse invariant mechanisms and finds a common flat minima, which in turn facilitates SIL. We further validate the effectiveness of SIL by demonstrating the generalization capabilities of ASGDRO on our newly developed synthetic SIL dataset, H-CMNIST, as well as on various types of distribution shift benchmark datasets.

6. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00457216)

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450, 2021. 3

[2] Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022. 4

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2, 3, 7

[4] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019. 7

[5] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34: 22405–22418, 2021. 2

[6] Liang Chen, Yong Zhang, Yibing Song, Zhiqiang Shen, and Lingqiao Liu. Lfme: A simple framework for learning from multiple experts in domain generalization. *arXiv preprint arXiv:2410.17020*, 2024. 7

[7] Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, Kaili Ma, Han Yang, Peilin Zhao, Bo Han, et al. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. *arXiv preprint arXiv:2206.07766*, 2022. 2

[8] Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 7, 8

[9] Ziliang Chen, Yongsen Zheng, Zhao-Rong Lai, Quanlong Guan, and Liang Lin. Diagnosing and rectifying fake ood invariance: A restructured causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11471–11479, 2024. 7

[10] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021. 3

[11] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. 6

[12] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016. 3

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020. 2, 5

[14] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 1, 2, 7, 8

[15] Siyuan Guo, Jonas Bernhard Wildberger, and Bernhard Schölkopf. Out-of-variable generalisation for discriminative models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[16] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. 4

[17] P Izmailov, AG Wilson, D Podoprikhin, D Vetrov, and T Garipov. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885, 2018. 1, 2

[18] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 2, 5, 6

[19] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 2, 3

[20] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 1, 2, 7

[21] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 3

[22] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021. 2, 5

[23] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 1

[24] Yong Lin, Lu Tan, Yifan Hao, Honam Wong, Hanze Dong, Weizhong Zhang, Yujiu Yang, and Tong Zhang. Spurious feature diversification improves out-of-distribution generalization. *arXiv preprint arXiv:2309.17230*, 2023. 2

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 7

[26] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013. 1

[27] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017. 2

[28] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020. 4

[29] Toan Nguyen, Kien Do, Bao Duong, and Thin Nguyen. Domain generalisation via risk distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2790–2799, 2024. 7

[30] Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019. 3

[31] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 2

[32] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022. 2

[33] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 28656–28679. PMLR, 2023. 4

[34] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. 2, 3

[35] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 1, 2, 3, 5, 6, 7

[36] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021. 2

[37] Jacob Mitchell Springer, Vaishnavh Nagarajan, and Aditi Raghunathan. Sharpness-aware minimization enhances feature quality via balanced learning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[38] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004. 1, 2

[39] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 2

[40] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 4

[41] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022. 1, 2, 7

[42] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020. 8

[43] Jianyu Zhang and Léon Bottou. Learning useful representations for shifting tasks and distributions. In *International Conference on Machine Learning*, pages 40830–40850. PMLR, 2023. 2

[44] Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich feature construction for the optimization-generalization dilemma. In *International Conference on Machine Learning*, pages 26397–26411. PMLR, 2022. 2, 3, 8

[45] Xin Zhang, Yusuke Iwasawa, Yutaka Matsuo, and Shixiang Shane Gu. Amortized prompt: Lightweight fine-tuning for clip in domain generalization. *arXiv preprint arXiv:2111.12853*, 2021. 8

[46] Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cu. Flatness-aware minimization for domain generalization. *arXiv preprint arXiv:2307.11108*, 2023. 2

[47] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. *arXiv preprint arXiv:2202.03599*, 2022. 4

[48] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C Dvornek, James s Duncan, Ting Liu, et al. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2021. 7

A Appendix: Sufficient Invariant Learning for Distribution Shift

A.1 Limitations and Future Works

ASGDRO utilizes adversarial perturbations to find flat minima, similar to SAM. It requires two forward and backward passes in a single training iteration, which is one of the persistent issues with SAM-based algorithms. However, recent research has been actively focusing on improving the computational cost of SAM (Du et al., 2021, 2022). The computational cost of ASGDRO can also be improved in a similar context, and we consider this to be a future work.

To evaluate whether the algorithm effectively learns diverse invariant mechanisms sufficiently and performs robust predictions, a new benchmark dataset is necessary. Unlike existing invariant learning benchmarks that only require a small number of attributes, constructing an SIL benchmark demands rich attribute annotations to form multiple invariant features. In this paper, we attempt to validate SIL using H-CMNIST, but it is a synthetic dataset based on MNIST. This implies the need for a new benchmark to validate SIL on real-world data, which we leave it as a future work.

A.2 The subset relationship of invariant features

In Definition 3, $h_{\theta_g}(\hat{Z}^{\mathrm{I}})$ refers to a classifier that relies solely on $\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}}$. Given a single sample, if any invariant feature within $\hat{Z}^{\mathrm{I}}$ is observed, we expect the loss evaluated by the classifier to be very small. For two different subset $\hat{Z}^{\mathrm{I}}_a, \hat{Z}^{\mathrm{I}}_b \subseteq \hat{Z}^{\mathrm{I}}$ that satisfy $\hat{Z}^{\mathrm{I}}_b \subseteq \hat{Z}^{\mathrm{I}}_a$, the following inequality holds:

$$P(\hat{Z}^{\mathrm{I}}_i \subseteq \hat{Z}^{\mathrm{I}}_b \text{ is observed in } e \in \mathcal{E}) \leq P(\hat{Z}^{\mathrm{I}}_i \subseteq \hat{Z}^{\mathrm{I}}_a \text{ is observed in } e \in \mathcal{E}).$$

where P denotes the probability. Note that Z^{I} also can be partitioned as follows:

$$Z^{\mathrm{I}} = \bigcup_{i=1}^{p} \{\hat{Z}^{\mathrm{I}} \mid |\hat{Z}^{\mathrm{I}}| = i\},$$

where $|\cdot|$ denotes the cardinality of a set and p the number of invariant features. It follows that

$$\begin{aligned}
\max_{\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}}} \mathbb{E}[\ell(h_{\theta_h}(\hat{Z}^{\mathrm{I}}), Y^e)] = \max \Big[& \mathbb{E}[\ell(h_{\theta_h}(Z^{\mathrm{I}}), Y^e)], \\
& \max_{\substack{\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}} \\ s.t. |\hat{Z}^{\mathrm{I}}| = p-1}} \mathbb{E}[\ell(h_{\theta_h}(\hat{Z}^{\mathrm{I}}), Y^e)], \\
& \dots, \\
& \max_{\substack{\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}} \\ s.t. |\hat{Z}^{\mathrm{I}}| = 1}} \mathbb{E}[\ell(h_{\theta_h}(\hat{Z}^{\mathrm{I}}), Y^e)] \Big] \\
= & \max_{\substack{\hat{Z}^{\mathrm{I}} \subseteq Z^{\mathrm{I}} \\ s.t. |\hat{Z}^{\mathrm{I}}| = 1}} \mathbb{E}[\ell(h_{\theta_h}(\hat{Z}^{\mathrm{I}}), Y^e)] \\
= & \max_{Z^{\mathrm{I}}_i \subset Z^{\mathrm{I}}} \mathbb{E}[\ell(h_{\theta_h}(Z^{\mathrm{I}}_i), Y^e)],
\end{aligned}$$

assuming that observing additional invariant features do not adversely affect the performance of the current model.

A.3 Proof of Proposition 1

Proposition 1. *By the Taylor expansion,*

$$\max_{e \in \mathcal{E}} \max_{||\epsilon_e|| \leq \rho} \mathcal{R}^e(\theta + \epsilon_e) \approx \max_{e \in \mathcal{E}} [\mathcal{R}^e(\theta) + \rho ||\nabla \mathcal{R}^e(\theta)||].$$

ASGDRO leads to a regularization of the gradient norm, $\mathcal{R}^e$, $||\nabla \mathcal{R}^e(\theta)||$, across environments, which drives the model to converge to common flat minima.

Proof. Recall that objective function of ASGDRO (Equation 9) is as follows:

$$\max_{e\in\mathcal{E}} \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta+\epsilon_e).$$

We use $\mathcal{E}$ instead of $\mathcal{E}_{\text{tr}}$, since this property of ASGDRO holds in any set of environments. As $\mathcal{R}^e(\theta)$ is independent of ϵ_e, it can be factored out of the maximization term over ϵ_e as follows:

$$\max_{e\in\mathcal{E}} \max_{||\epsilon_e||\leq\rho} \mathcal{R}^e(\theta+\epsilon_e) = \max_{e\in\mathcal{E}}[\mathcal{R}^e(\theta) + \max_{||\epsilon_e||\leq\rho}[\mathcal{R}^e(\theta+\epsilon_e) - \mathcal{R}^e(\theta)]]$$

Note that we intentionally add and subtract $\mathcal{R}_e$ to reformulate the expression, enabling the separation of terms for clearer analysis. Using the Taylor approximation expanded up to the first-order term, we have:

$$\begin{aligned}\max_{e\in\mathcal{E}}[\mathcal{R}^e(\theta) + \max_{||\epsilon_e||\leq\rho}[\mathcal{R}^e(\theta+\epsilon_e) - \mathcal{R}^e(\theta)]] &\approx \max_{e\in\mathcal{E}}[\mathcal{R}^e(\theta) + \max_{||\epsilon_e||\leq\rho}[\epsilon_e\cdot\nabla\mathcal{R}^e(\theta)]] \\ &= \max_{e\in\mathcal{E}}[\mathcal{R}^e(\theta) + \epsilon_e^*\cdot\nabla\mathcal{R}^e(\theta)], \end{aligned} \tag{5}$$

where $\epsilon_e^* = \rho\frac{\nabla\mathcal{R}^e(\theta)}{||\nabla\mathcal{R}^e(\theta)||}$. Note that Eq. 5 holds because the maximum value over $||\epsilon_e|| \leq \rho$ is achieved when ϵ_e and $\nabla\mathcal{R}^e(\theta)$ are aligned in the same direction (Foret et al., 2020). By substituting ϵ_e^*, we obtain the following equation:

$$\max_{e\in\mathcal{E}}[\mathcal{R}^e(\theta) + \epsilon_e^*\cdot\nabla\mathcal{R}^e(\theta)] = \max_{e\in\mathcal{E}}[\mathcal{R}^e(\theta) + \rho||\nabla\mathcal{R}^e(\theta)||]. \tag{6}$$

Zhao et al. (2022) demonstrate that minimizing the gradient norm of the risk leads to finding flat minima. Eq. 6 minimizes both risk and the gradient norm of risk for each environment. Consequently, ASGDRO constrains the training process to find a common flat minimum across environments.

□

A.4 Proof of Theorem 1

Theorem 1. *Let $\theta_\lambda^{\text{I}}$ be a convex combination of θ_i^{I}, where λ is a p-dimensional vector. Consider mean-squared error as the loss function. Assume a linear model with $Z \in \mathbb{R}^p$, where the p features are orthogonal, and suppose $Z = Z^{\text{I}} = (1,\ldots,1)$. Then,*

$$\begin{aligned}\lambda^* &= \operatorname*{argmin}_\lambda \max_{e\in\mathcal{E}_{tr}} \max_{||\epsilon||\leq\rho} \mathcal{R}^e(\theta_\lambda^{\text{I}} + \epsilon) \\ &\approx \operatorname*{argmin}_\lambda \max_{e\in\mathcal{E}_{tr}} \left[\mathcal{R}^e(\theta_\lambda^{\text{I}}) + \rho||\lambda||\cdot||\nabla\mathcal{R}^e(\theta_\lambda^{\text{I}})||\right] \\ &= \operatorname*{argmin}_\lambda ||\lambda|| = (\frac{1}{p},\ldots,\frac{1}{p})\end{aligned} \tag{7}$$

where $||\cdot||$ denotes L_2 norm.

Proof. In this setting, we consider a single input for each environment e. Suppose there are p invariant features, and every invariant feature has the same activation:

$$Z^{\text{I}} = (1,\ldots,1),$$

where $|Z^{\text{I}}| = p$. We assume that all spurious features are completely removed. Thus, $Z = (Z^{\text{I}}, Z^{\text{NI}}) = Z^{\text{I}}$, where $|Z| = p$. Consequently, the risk for Z is identical across all environments e:

$$\mathcal{R}^e(\theta) = \mathcal{R}^{e'}(\theta) = c \quad \text{for any } e, e' \in \mathcal{E}_{\text{tr}}, \tag{8}$$

where c is a constant. Given Z^{I}, we focus only on the parameters of the classifier, denoted by θ^{I}. Recall that the classifier satisfying Eq. 3 in main paper, and Eq. 8, is not unique. Define θ_i^{I} as the classifier that utilizes only the i-th element of Z^{I}.

For simplicity, let θ_i^{I} be a column vector where only the i–th element is one, and all other elements are zero:

$$Z^{\text{I}}\theta_i^{\text{I}} = Z_i^{\text{I}} = 1. \tag{9}$$

Furthermore, the convex combination of θ_i^{I} also yields an equivalent output:

$$Z^{\mathrm{I}} \sum_{i=1}^{p} \lambda_i \theta_i^{\mathrm{I}} = 1,$$

where $\sum_{i=1}^{p} \lambda_i = 1$ and $0 \leq \lambda_i \leq 1$ for all $i \in \{1, \ldots, p\}$. We denote the current classifier as $\theta_\lambda^{\mathrm{I}} := \sum_{i=1}^{p} \lambda_i \theta_i^{\mathrm{I}}$, where $\lambda = (\lambda_1, \ldots, \lambda_p)$. From Proposition 1, we know:

$$\max_{e \in \mathcal{E}} \max_{||\epsilon_e|| \leq \rho} \mathcal{R}^e(\theta + \epsilon_e) = \max_{e \in \mathcal{E}_{\mathrm{tr}}} \left[\mathcal{R}^e(\theta) + \rho ||\nabla_\theta \mathcal{R}^e(\theta)|| \right]. \tag{10}$$

For the mean-squared error loss function $\mathcal{R}^e(\theta) = \frac{1}{2}\|Y^e - \sum_{i=1}^{p} \theta_i\|^2$, the gradient is given by $\nabla \mathcal{R}^e(\theta) = -(Y^e - \sum_{i=1}^{p} \theta_i) \cdot \mathbf{1}$, where $\mathbf{1}$ is a p-dimensional vector whose elements are all equal to 1. Substituting $\theta_\lambda^{\mathrm{I}}$ into Eq. 10, we get:

$$\max_{e \in \mathcal{E}} \max_{\|\epsilon_e\| \leq \rho} \mathcal{R}^e(\theta_\lambda^{\mathrm{I}} + \epsilon_e) = \max_{e \in \mathcal{E}_{\mathrm{tr}}} \left[\mathcal{R}^e(\theta_\lambda^{\mathrm{I}}) + \rho \|\nabla_\theta \mathcal{R}^e(\theta_\lambda^{\mathrm{I}})\| \right].$$

This simplifies to:

$$\max_{e \in \mathcal{E}_{\mathrm{tr}}} \left[\mathcal{R}^e(\theta_\lambda^{\mathrm{I}}) + \rho \| - \lambda \odot \nabla \mathcal{R}^e(\theta_\lambda^{\mathrm{I}})\| \right] = \max_{e \in \mathcal{E}_{\mathrm{tr}}} \left[\mathcal{R}^e(\theta_\lambda^{\mathrm{I}}) + \rho \|\lambda\| \cdot \|\nabla \mathcal{R}^e(\theta_\lambda^{\mathrm{I}})\| \right],$$

where $\mathcal{R}^e(\theta_\lambda^{\mathrm{I}}) = c$ for any λ. Since the classifier uses only invariant features, minimizing the adversarial term reduces to:

$$\begin{aligned} \operatorname*{argmin}_{\lambda} \max_{e \in \mathcal{E}_{\mathrm{tr}}} \max_{||\epsilon|| \leq \rho} \mathcal{R}^e(\theta_\lambda^{\mathrm{I}} + \epsilon) &= \operatorname*{argmin}_{\lambda} \max_{e \in \mathcal{E}_{\mathrm{tr}}} \left[\mathcal{R}^e(\theta_\lambda^{I}) + \rho ||\lambda|| \cdot ||\nabla \mathcal{R}^e(\theta_\lambda^{I})|| \right] \\ &= \operatorname*{argmin}_{\lambda} ||\lambda||. \end{aligned}$$

By the Cauchy-Schwarz inequality:

$$\left(\sum_{i=1}^{p} \lambda_i \right)^2 \leq p \cdot \sum_{i=1}^{p} \lambda_i^2 = p \cdot ||\lambda||^2.$$

Under the condition $\sum_{i=1}^{p} \lambda_i = 1$, equality holds when $\lambda_i = \frac{1}{p}$ for all i, yielding:

$$\operatorname*{argmin}_{\lambda} ||\lambda|| = (\frac{1}{p}, \ldots, \frac{1}{p})$$

□

A.5 Mechanism of ASGDRO for Removing Spurious Features

ASGDRO successfully removes spurious features. Inspired by Andriushchenko et al. (2023), we reformulate the two-layer ReLU case presented in that paper to demonstrate this. Consider a two–layer ReLU network

$$f(\theta) = \langle \theta_h, \sigma(\theta_g x) \rangle,$$

where $\theta = (\theta_g, \theta_h)$, $\theta_g \in \mathbb{R}^{k \times m}$ and $\theta_h \in \mathbb{R}^k$. Recall that ASGDRO minimizes the maximum sharpness across environments:

$$\max_{e \in \mathcal{E}} \max_{||\epsilon_e|| \leq \rho} \mathcal{R}^e(\theta + \epsilon_e).$$

Let e_t denote the environment that attains the maximum risk at the current step t. Then, the adversarial perturbation is $\epsilon_{e_t}^* = \rho \frac{\nabla \mathcal{R}^{e_t}(\theta)}{\|\nabla \mathcal{R}^{e_t}(\theta)\|}$ (Foret et al., 2020) and the risk is

$$\max_{||\epsilon_{e_t}|| \leq \rho} \mathcal{R}^{e_t}(\theta + \epsilon_{e_t}) = \mathcal{R}^{e_t}(\theta + \rho \frac{\nabla \mathcal{R}^{e_t}(\theta)}{\|\nabla \mathcal{R}^{e_t}(\theta)\|})$$

Under the first–order Taylor approximation,

$$\nabla \mathcal{R}^{e_t} \left(\theta + \rho \frac{\nabla \mathcal{R}^{e_t}(\theta)}{\|\nabla \mathcal{R}^{e_t}(\theta)\|} \right) \approx \nabla \left[\mathcal{R}^{e_t}(\theta) + \rho \|\nabla \mathcal{R}^{e_t}(\theta)\| \right]$$

Andriushchenko et al. (2023) shows that under two–layer ReLU network, the update rule for pre-activation of k–th neuron is as follows:

$$\langle \theta_g^{(k)}, x \rangle^{(t+1)} \approx \langle \theta_g^{(k)}, x \rangle^{(t)} \underbrace{-\eta\gamma \left(1 + \rho \frac{\|\nabla f(\theta)\|}{\sqrt{\mathcal{R}^{e_t}(\theta)}}\right) a_k \sigma'(\langle \theta_g^{(k)}, x \rangle) \|x\|^2}_{\text{(a)}}$$
$$\underbrace{-\eta\rho \frac{\sqrt{\mathcal{R}^{e_t}(\theta)}}{\|\nabla f(\theta)\|} \sigma(\langle \theta_g^{(k)}, x \rangle) \|x\|^2}_{\text{(b)}},$$

where η denotes the learning rate, $\gamma = f(\theta) - y$, i.e. the residual.

In ASGDRO, regularization on the gradient norm has two key effects. First, as seen in term (a), the gradient update direction remains the same, but the model is updated with a larger learning rate. Second, in term (b), when $\mathcal{R}^{e_t}(\theta)$ is large enough, the pre-activation of the k-th neuron, $\langle \theta_g^{(k)}, x \rangle$, turns negative. Note that a large $\mathcal{R}^{e_t}$ implies that highly activated neurons at this point tend to encode significant information from spurious features. When $\mathcal{R}^{e_t}(\theta)$ causes the pre-activation of a neuron to become negative, the nature of the ReLU activation function ensures that the output of that neuron becomes zero. This indicates that, under distribution shifts, regularization via the common flat minima in ASGDRO effectively removes spurious features.

A.6 Heterogeneous-CMNIST (H-CMNIST)

Dataset Details

The test set of H-CMNIST is constructed by flipping the proportion of Z_{BP} from the training set. H-CMNIST conducts two types of tests. First, TestBed 1 evaluates whether the algorithm learns at least one invariant feature. To assess this, it compares the prediction differences between cases where the spurious feature Z_{BP} is present and absent. TestBed 2 examines whether the model remains robust to Z_{BP} and maintains good performance when only Z_{shape} is present, excluding Z_{color} among the two invariant features.

Experimental Details

In H-CMNIST experiments, we use ResNet18 (He et al., 2016) with SGD. We also conduct reweighted sampling when the algorithm setting can use the environment information, i.e., GDRO (Sagawa et al., 2019) and ASGDRO. In the H-CMNIST experiment, we set the loss of GDRO and ASGDRO by the group, not the domain. That is, there is four groups; (Class=0,BP=Top Left), (Class=0,BP=Bottom Right), (Class=1,BP=Top Left), (Class=1,BP=Bottom Right). For hyperparameter tuning, we perform grid search over learning rate, $\{10^{-3}, 10^{-4}\}$, and L_2–regularization, $\{1, 10^{-1}, 10^{-3}, 10^{-4}\}$. We fix the batch size, 128, and train the model up to 20 epochs. For ASAM (Kwon et al., 2021) and ASGDRO, we search the hyperparameter ρ among $\{0.05, 0.2, 0.5, 0.8\}$. We fix the robust step size, γ, as 0.01 for GDRO and ASGDRO. We evaluate the models with three random seeds.

A.7 Subpopulation Shifts: Datasets and Experimental Details

Dataset Details

In Table 2 in the main paper, we conduct our experiment for subpopulation shifts with five datasets: CMNIST (Arjovsky et al., 2019), Waterbirds (Sagawa et al., 2019), CelebA (Liu et al., 2015), CivilComments (Borkan et al., 2019). CMNIST, Waterbirds, and CelebA datasets correspond to computer vision tasks (Figure 6), while CivilComments pertain to natural language processing tasks. In this section, we will describe each dataset and provide experimental details. To implement this, we utilized the codes provided by (Yao et al., 2022)[1].

[1] https://github.com/huaxiuyao/LISA

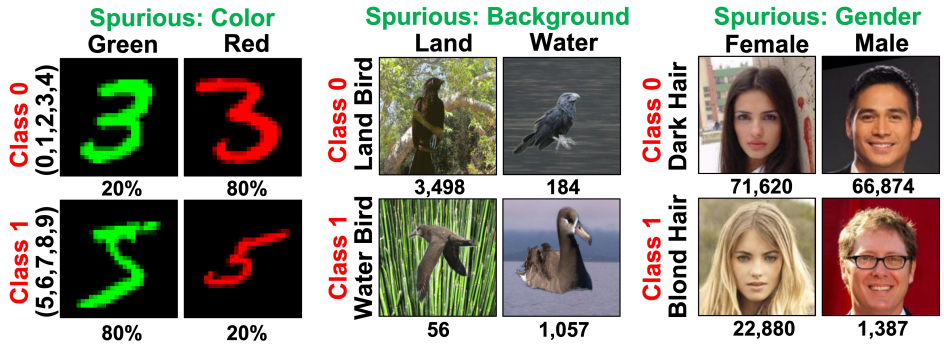

Figure 6: **CMNIST, Waterbirds, CelebA.** In each dataset, each row represents the class and each column represents the spurious feature. The numbers written below the images represent the ratio or count of data belonging to each group in the training dataset, where each group consists of (Class, Spurious Feature) pairs.

Colored MNIST (CMNIST)

In the CMNIST dataset provided by (Arjovsky et al., 2019), we perform binary classification to predict which number corresponds to the shape of a given digit. Specifically, when the shape of the digit corresponds to a logit between 0 and 4, the class is assigned as 0, and when it falls between 5 and 9, the class is assigned as 1. However, unlike the original MNIST dataset (LeCun et al., 1998), CMNIST introduces color as a spurious feature in the training set. When this spurious correlation becomes stronger than the invariant relationship between the class and the shape of the digit, a model trained without any regularization may be prone to relying on the spurious feature for predictions.

While Arjovsky et al. (2019) constructs two environments with different ratios of spurious features in the training set, Yao et al. (2022) uses a single environment to compose the training set. Our CMNIST dataset experiment follows the same setting as (Yao et al., 2022), where the dataset consists of four groups when considering combinations of "Shape of Logit" and "Color" as a single group. Concretely, Class 0 and Class 1 have similar numbers of data points, but the distribution of spurious features differs between the two classes. Class 0 consists of 80% red logits and 20% green logits, while Class 1 has 80% green logits and 20% red logits. Furthermore, within each class, 25% of the data acts as label noise, having a logit shape that does not correspond to its class. Therefore, the spurious feature, color, forms a stronger correlation between classes compared to that of the invariant feature, the shape of logits.

The validation set is constructed with an equal number of instances per group. The worst-group accuracy, defined as the lowest accuracy among all the groups, is utilized to select the best model. For the test set, we assume a distribution of the spurious feature that is opposite to the training set. Specifically, for Class 0, 90% of the data has a red color, and 10% has a green color, while for Class 1, it is the opposite. It is done to assess whether the model relies on the spurious feature for predictions.

Waterbirds

Waterbirds dataset, constructed by (Sagawa et al., 2019), is designed for the task of determining whether a bird belongs to the Landbird or Waterbird class. It consists of images of birds, from (Wah et al., 2011), as the invariant feature, while the spurious feature is the background, from (Zhou et al., 2017), which can either be Water or Land background. Indeed, in the Waterbirds dataset, the groups are formed by the combination of "Bird" and "Background". Specifically, the bird images corresponding to each class consist of more than 10 different species of birds. On the other hand, each background is composed of two categories obtained from (Zhou et al., 2017). As can be seen in Figure 6, the Landbird class predominantly has images with Land background, while the majority of images in the Waterbird class have Water background. Therefore, the spurious feature, background, may indeed form a strong spurious correlation with each class.

We follow the setting of previous research, (Sagawa et al., 2019; Yao et al., 2022), for the validation and test processes as well. The best model is selected based on the highest worst-group accuracy on the validation set. Unlike the training set, the validation and test sets are designed to have an equal

number of images for each group within each class. When reporting the average accuracy on the test dataset using the best model, we first compute the group accuracy for each group in the test set. Then, we calculate the weighted average of these accuracies using the group distribution from the training set. This approach is adopted to mitigate the uncertainty in estimating group accuracies, as the number of images belonging to the minority group in the Waterbird dataset is significantly smaller compared to other datasets (Sagawa et al., 2019).

CelebA

CelebA dataset by (Liu et al., 2015) is a collection of facial images of celebrities from around the world. It includes attribute values associated with each individual, such as hair color and gender. In order to evaluate the effects of subpopulation shifts, Sagawa et al. (2019) reformulated the CelebA dataset to align with the task of predicting whether the hair color is blond or not. In this case, the spurious feature is gender, and thus, the dataset is composed of four groups based on the combinations of hair color and gender. It can be observed from Figure 6 that images belonging to Class 0, corresponding to dark hair rather, are plentiful regardless of gender. However, for images in Class 1, which represent blond hair, the majority of them are distributed in the Female group. Therefore, gender can act as a spurious feature, and the goal of this task is to obtain a model that focuses solely on the invariant feature, hair color, rather than the face which may capture the characteristics of gender-related features.

The best model is selected based on the best worst-group accuracy on the validation set. In this case, the validation set and test set have the same distribution of images per group as the training set. Therefore, the average test accuracy reflects this distribution accordingly.

CivilComments

The CivilComments dataset, (Borkan et al., 2019), is a dataset that gathers comments from online platforms and is used for the task of classifying whether a given comment is toxic or not. We conduct the experiment on the CivilComments dataset, which has been reformulated by (Koh et al., 2021). Each comment is labeled to indicate whether it mentions the presence of any word of the 8 demographic identities; Black, White, Christian, Muslim, other religions, Male and Female. Therefore, the CivilComments dataset consists of 16 groups, formed by the combination of toxic labels and the presence or absence of the 8 demographic identities in each comment. Each demographic identity can potentially act as a spurious feature. To prevent this, the goal of the task is to train the model to focus solely on the invariant feature of toxic labels and not rely on demographic identities as predictive factors.

However, in reality, unlike other datasets, each comment in the CivilComments dataset can mention more than one demographic identity. Considering all possible combinations of demographic identities for each comment and training the model on all these combinations would be inefficient. Therefore, we follow the learning approach proposed by (Koh et al., 2021). Concretely, we only consider four groups based on whether the comment mentions toxicity and whether it mentions the demographic identity of being "Black", without considering other demographic identities. We train the model using these four groups. However, during the validation and test, we evaluate the model's performance individually for all 16 groups and record the lowest accuracy among the group accuracies as the worst-group accuracy. The Best model is selected based on this worst-group accuracy.

Experimental Details

The search range of the hyperparameter ρ, which determines the range for exploring the flat region, is fixed to $\{0.05, 0.2, 0.5, 0.8, 1.0, 1.2, 1.5\}$ for all datasets. We evaluate the model across three random seeds and report the average performance. We set robust step size γ, in Algorithm 1 of the main paper, $\{0.1, 0.01\}$. In addition, we use the same range for adjusted-group coefficient C, $\{0, 1, 2, 3, 4, 5\}$ (Section 3.3 in (Sagawa et al., 2019) for details). In CMNIST, Waterbirds, and CelebA datasets, we utilize ResNet50 (He et al., 2016) models. The same hyperparameter ranges are applied to ASAM and ASGDRO, and the other performances for other baselines are reported performances from (Liu et al., 2021; Yao et al., 2022; Han et al., 2022). All experiments in this paper were conducted using NVIDIA RTX A6000 with 49140 MiB of GPU memory and GeForce RTX 3090 with 24.00 GiB of GPU memory.

	CMNIST		Waterbirds		CelebA		CivilComments	
	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
ERM‡	27.8± 1.9%	0.0± 0.0%	97.0± 0.2%	63.7± 1.9%	**94.9± 0.2%**	47.8± 3.7%	92.2± 0.1%	56.0± 3.6%
ASAM	40.5± 0.8%	34.1± 1.2%	**97.4± 0.0%**	72.4± 0.4%	93.7± 0.8%	46.5± 10.3%	**92.3± 0.1%**	58.9± 1.7%
IRM‡	72.1± 1.2%	70.3± 0.8%	87.5± 0.7%	75.6± 3.1%	94.0± 0.4%	77.8± 3.9%	88.8± 0.7%	66.3± 2.1%
IB-IRM‡	72.2± 1.3%	70.7± 1.2%	88.5± 0.6%	76.5± 1.2%	93.6± 0.3%	85.0± 1.8%	89.1± 0.3%	65.3± 1.5%
V-REx‡	71.7± 1.2%	70.2± 0.9%	88.0± 1.0%	73.6± 0.2%	92.2± 0.1%	86.7± 1.0%	90.2± 0.3%	64.9± 1.2%
CORAL‡	71.8± 1.7%	69.5± 0.9%	90.3± 1.1%	79.8± 1.8%	93.8± 0.3%	76.9± 3.6%	88.7± 0.5%	65.6± 1.3%
GDRO‡	72.3± 1.2%	68.6± 0.8%	91.8± 0.3%	90.6± 1.1%	92.1± 0.4%	87.2± 1.6%	89.9± 0.5%	70.0± 2.0%
DomainMix‡	51.4± 1.3%	48.0± 1.3%	76.4± 0.3%	53.0± 1.3%	93.4± 0.1%	65.6± 1.7%	90.9 ± 0.4%	63.6± 2.5%
Fish‡	46.9± 1.4%	35.6± 1.7%	85.6± 0.4%	64.0± 0.3%	93.1± 0.3%	61.2± 2.5%	89.8± 0.4%	71.1± 0.4%
LISA‡	74.0± 0.1%	73.3± 0.2%	91.8± 0.3%	89.2± 0.6%	92.4± 0.4%	89.3± 1.1%	89.2± 0.9%	**72.6± 0.1%**
PDE‡‡	–%	–%	92.4± 0.8%	90.3± 0.3%	92.0± 0.6%	**91.0± 0.4%**	86.3± 1.7%	71.5± 0.5%
ASGDRO	**74.8± 0.1%**	**74.2± 0.0%**	92.3± 0.1%	**91.4± 0.1%**	92.1± 0.4%	**91.0± 0.5%**	90.2± 0.2%	71.8± 0.4%

Table 5: **Subpopulation Shift**. ‡ denotes the performance reported from (Yao et al., 2022), and ‡‡ denotes the performance reported from (Deng et al., 2024). Avg. denotes average accuracy, and Worst denotes worst group accuracy

In CMNIST, we have the same hyperparameter search range as (Yao et al., 2022) by default: batch size 16, learning rate 10^{-3}, L_2–regularization 10^{-4} with SGD over 300 epochs. For Waterbirds, we perform the grid search over the batch size, $\{16, 64\}$, the learning rate, $\{10^{-3}, 10^{-4}, 10^{-5}\}$, and L_2–regularization, $\{10^{-4}, 10^{-1}, 1\}$. We train our model with SGD over 300 epochs. We also conduct grid search over the batch size, $\{16, 128\}$, the learning rate, $\{10^{-4}, 10^{-5}\}$, and L_2–regularization, $\{10^{-4}, 10^{-2}, 1\}$ for CelebA, training with SGD over 50 epochs. We referenced (Yao et al., 2022; Liu et al., 2021) for this range of hyperparameter search. For CivilComments, we use DistilBERT (Sanh et al., 2019) model. We follow the hyperparameter search range provided in (Koh et al., 2021). For optimizer, we use AdamW (Loshchilov and Hutter, 2017) with 10^{-2} for L_2–regularization. We find the optimal learning rate among $\{10^{-6}, 2 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}\}$. We train up to 5 epochs with batch size 16. The gradient clipping is applied only during the second step, which is the actual update step, in the SAM-based algorithm (Foret et al., 2020).

A.8 Error bars for Wilds Benchmark

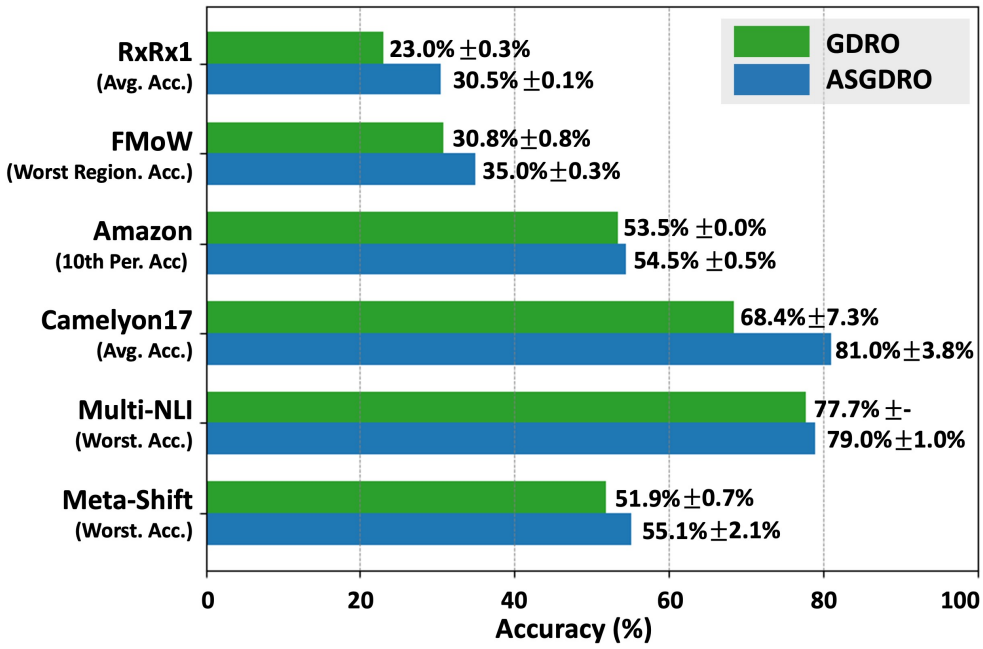

Figure 7: **Standard Deviations for Wilds Benchmark Datasets.**

We demonstrate the differences between GDRO and ASGDRO in various distribution shift scenarios that could occur in the real world. Wilds benchmark Koh et al. (2021) consists of datasets collected

from the real world. Camelyon17 and RxRx1 are datasets where domain shift is predominant. Amazon and FMoW are datasets where both subpopulation shift and domain shift are simultaneously predominant. Figure 7 shows the results of ASGDRO and GDRO on Wilds Benchmark, MetaShift dataset, and Multi-NLI (Williams et al., 2017). ASGDRO shows superior performances consistently compared with GDRO. It implies that identifying common flat minima across environments enhances the robustness of models.

A.9 Experimental Details and Error bars for Domainbed with DPLCLIP

Experimental Details for DomainBed Experiment

Using DomainBed framework (Gulrajani and Lopez-Paz, 2020), we evaluate domain generalization algorithms by randomly sampling hyperparameter combinations within predefined hyperparameter search ranges for each algorithm. The goal of domain generalization is to train models that perform robustly on unseen domains. Consequently, the choice of the best model is heavily influenced by whether the validation set used for model selection is sampled from the test domain or the train domains. To account for this, we provide results for both the training-domain validation set, which does not utilize information from the test domain, and the test-domain validation set, where model selection is performed using information from the test domain. The following subsections present the results for each dataset, considering both model selection methods.

By combining ASGDRO with the existing successful domain generalization approach, DPLCLIP (Zhang et al., 2021)[2], we demonstrate the versatility of ASGDRO, as it can easily be integrated with other algorithms. Moreover, our results show that ASGDRO not only improves performance in the context of subpopulation shift but also achieves performance gains in the presence of domain shift. For experimental details, we set the range of the robust step size γ as `lambda r: 10**r.uniform(-4, -2)` with $\gamma = 0.001$ by default and the neighborhood size ρ as `lambda r: r.choice([0.05, 0.5, 1.0, 5.0])`. The other settings are the same as DPLCLIP (Zhang et al., 2021). Following common convention, we conduct 20 hyperparameter searches and reported the averages for three random seeds. We evaluated our model on the five datasets: VLCS (Fang et al., 2013), PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), TerraIncognita (Beery et al., 2018) and DomainNet (Peng et al., 2019). For the original ASGDRO experiments, we follow the experimental setup of Wang et al. (2023)[3].

Model selection: training-domain validation set

VLCS

Algorithm	C	L	S	V	Avg
DPLCLIP	99.1 ± 0.5	61.1 ± 1.5	72.6 ± 2.6	83.1 ± 2.5	79.0
DPLCLIP GDRO	99.9 ± 0.0	61.3 ± 2.5	74.4 ± 1.1	83.4 ± 2.6	79.7
DPLCLIP ASGDRO	100.0 ± 0.0	62.7 ± 0.4	74.5 ± 1.4	85.7 ± 0.8	80.7

PACS

Algorithm	A	C	P	S	Avg
DPLCLIP	97.6 ± 0.2	98.3 ± 0.3	99.9 ± 0.0	90.5 ± 0.5	96.6
DPLCLIP GDRO	97.0 ± 0.7	98.2 ± 0.1	99.8 ± 0.1	88.6 ± 1.4	95.9
DPLCLIP ASGDRO	97.7 ± 0.1	98.7 ± 0.1	99.8 ± 0.0	91.0 ± 0.5	96.8

OfficeHome

Algorithm	A	C	P	R	Avg
DPLCLIP	80.6 ± 0.8	69.2 ± 0.2	90.1 ± 0.2	91.1 ± 0.0	82.7
DPLCLIP GDRO	82.3 ± 0.2	70.9 ± 0.1	90.0 ± 0.4	91.1 ± 0.1	83.6
DPLCLIP ASGDRO	82.1 ± 0.4	71.3 ± 0.8	90.3 ± 0.6	91.2 ± 0.3	83.7

[2]https://github.com/shogi880/DPLCLIP
[3]https://github.com/Wang-pengfei/SAGM

TerraIncognita

Algorithm	L100	L38	L43	L46	Avg
DPLCLIP	47.1 ± 1.4	50.1 ± 1.2	41.6 ± 1.9	42.7 ± 0.7	45.4
DPLCLIP GDRO	49.1 ± 0.9	48.7 ± 2.6	46.3 ± 2.6	39.8 ± 1.4	46.0
DPLCLIP ASGDRO	52.8 ± 0.9	51.5 ± 2.1	49.2 ± 1.2	42.1 ± 0.9	48.9

DomainNet

Algorithm	clip	info	paint	quick	real	sketch	Avg
DPLCLIP	70.9 ± 0.3	51.9 ± 0.3	66.6 ± 0.3	14.6 ± 0.5	84.3 ± 0.2	66.6 ± 0.1	59.1
DPLCLIP GDRO	71.8 ± 0.4	51.3 ± 0.4	67.0 ± 0.3	15.3 ± 0.2	84.4 ± 0.1	65.0 ± 0.9	59.1
DPLCLIP ASGDRO	71.5 ± 0.5	52.2 ± 0.4	67.5 ± 0.6	16.4 ± 0.2	84.7 ± 0.1	66.5 ± 0.2	59.8

Averages

Algorithm	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet	Avg
DPLCLIP	79.0 ± 0.7	96.6 ± 0.1	82.7 ± 0.2	45.4 ± 1.0	59.1 ± 0.1	72.6
DPLCLIP GDRO	79.7 ± 1.3	95.9 ± 0.4	83.6 ± 0.1	46.0 ± 1.0	59.1 ± 0.2	72.9
DPLCLIP ASGDRO	80.7 ± 0.3	96.8 ± 0.2	83.7 ± 0.5	48.9 ± 0.3	59.8 ± 0.2	74.0

Model selection: test-domain validation set (Oracle)

VLCS

Algorithm	C	L	S	V	Avg
DPLCLIP	99.8 ± 0.1	69.7 ± 0.6	72.4 ± 1.0	86.2 ± 0.5	82.0
DPLCLIP GDRO	99.9 ± 0.0	64.9 ± 1.1	79.1 ± 0.5	86.5 ± 0.2	82.6
DPLCLIP ASGDRO	99.8 ± 0.1	67.4 ± 0.9	78.1 ± 0.5	86.9 ± 0.1	83.1

PACS

Algorithm	A	C	P	S	Avg
DPLCLIP	97.6 ± 0.1	98.7 ± 0.3	99.8 ± 0.1	91.2 ± 0.3	96.8
DPLCLIP GDRO	97.4 ± 0.3	98.9 ± 0.2	99.8 ± 0.1	91.9 ± 0.3	97.0
DPLCLIP ASGDRO	97.7 ± 0.2	99.1 ± 0.0	99.9 ± 0.0	91.7 ± 0.3	97.1

OfficeHome

Algorithm	A	C	P	R	Avg
DPLCLIP	81.7 ± 0.2	70.9 ± 0.1	90.3 ± 0.3	90.7 ± 0.0	83.4
DPLCLIP GDRO	81.3 ± 0.8	70.6 ± 0.3	90.5 ± 0.1	90.9 ± 0.3	83.3
DPLCLIP ASGDRO	83.2 ± 0.4	71.7 ± 0.2	91.9 ± 0.1	91.3 ± 0.1	84.5

TerraIncognita

Algorithm	L100	L38	L43	L46	Avg
DPLCLIP	55.9 ± 2.3	58.5 ± 0.3	48.2 ± 0.5	40.9 ± 3.0	50.9
DPLCLIP GDRO	57.9 ± 1.0	55.3 ± 1.5	49.6 ± 2.0	41.8 ± 1.4	51.2
DPLCLIP ASGDRO	56.2 ± 0.8	54.1 ± 0.3	50.7 ± 0.7	42.1 ± 0.5	50.8

DomainNet

Algorithm	clip	info	paint	quick	real	sketch	Avg
DPLCLIP	72.0 ± 0.5	52.1 ± 0.3	67.3 ± 0.2	16.6 ± 0.2	84.4 ± 0.2	66.8 ± 0.1	59.9
DPLCLIP GDRO	72.0 ± 0.2	51.7 ± 0.1	67.2 ± 0.4	16.7 ± 0.2	84.5 ± 0.0	66.3 ± 0.1	59.7
DPLCLIP ASGDRO	71.5 ± 0.5	52.8 ± 0.3	68.1 ± 0.3	16.5 ± 0.2	84.9 ± 0.0	67.0 ± 0.1	60.2

Averages

Algorithm	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet	Avg
DPLCLIP	82.0 ± 0.3	96.8 ± 0.1	83.4 ± 0.1	50.9 ± 0.6	59.9 ± 0.2	74.6
DPLCLIP GDRO	82.6 ± 0.2	97.0 ± 0.2	83.3 ± 0.2	51.2 ± 1.0	59.7 ± 0.0	74.8
DPLCLIP ASGDRO	83.1 ± 0.2	97.1 ± 0.1	84.5 ± 0.1	50.8 ± 0.3	60.2 ± 0.1	75.1

A.10 Grad-CAM Analysis

In this section, we present additional Grad-CAM (Selvaraju et al., 2017) results on the Waterbirds and CelebA datasets. In Figure 8 and 9, the red-colored-name features represent invariant features in the respective task, while the green-colored-name features represent spurious features. In the Grad-CAM images, the pixels that each model focuses on to predict the ground-truth label are highlighted closer to the red color in the image.

ERM (Vapnik, 1999) and ASAM (Kwon et al., 2021) are regularization-free algorithms that do not specifically encourage models to focus on invariant features, and this is reflected in the Grad-CAM results. Specifically, when observing Group 0 and Group 3 of Waterbirds, which can strongly form the correlation between class and spurious, as well as Group 0, 1, and 2 of CelebA, in most cases, the results show a strong focus on both spurious and invariant features simultaneously or solely on spurious features. For some images, particularly between CelebA dataset's Group 0 and 1 where there are no minority groups within a class, there is some degree of focus on invariant features. However, these images still contain a significant amount of unnecessary pixels such as the background. Conversely, in minority groups such as Group 1 and 2 in Waterbirds or Group 3 in CelebA, there is a predominant focus on invariant features to predict the ground-truth label. However, this focus is limited to only a subset of the overall invariant features and still include some spurious features.

In algorithms specifically designed to learn invariant features like GDRO (Sagawa et al., 2019), LISA (Yao et al., 2022), and ASGDRO (Ours), the Grad-CAM results exhibit different patterns compared to ERM and ASAM. In the most of results for the three algorithms, the models demonstrate a reasonable focus on invariant features. Compared with ERM and ASAM, there are significant reductions in the extent to which they focus on spurious features. However, GDRO and LISA still concentrate only on a part of invariant features. Additionally, in some cases, they may exhibit a greater focus on spurious features than on the subset of invariant features. It is also frequently observed that they still heavily include spurious features or solely focus on spurious features when dealing with majority groups such as Group 1 and 3 in Waterbirds or Group 0, 1, and 2 in CelebA. As in the results of Group 1, and 2 in Waterbirds or Group 3 in CelebA, we observe that the models' low ability to fully concentrate on invariant features is affected by the performance of models that still exhibit a focus on spurious features. This observation highlights the impact of the models' performance on their ability to completely focus on invariant features.

In contrast to other baselines, ASGDRO demonstrates a stronger focus on invariant features. As a result, Grad-CAM analysis shows that ASGDRO has relatively larger regions of focus on invariant features compared to other baselines. Simultaneously, it successfully eliminates spurious features while accurately predicting the ground-truth label. Therefore, these results demonstrate that ASGDRO has a higher capacity for capturing sufficiently diverse invariant features, and this characteristic is reflected in its performance. That is, ASGDRO promotes that the model performs SIL.

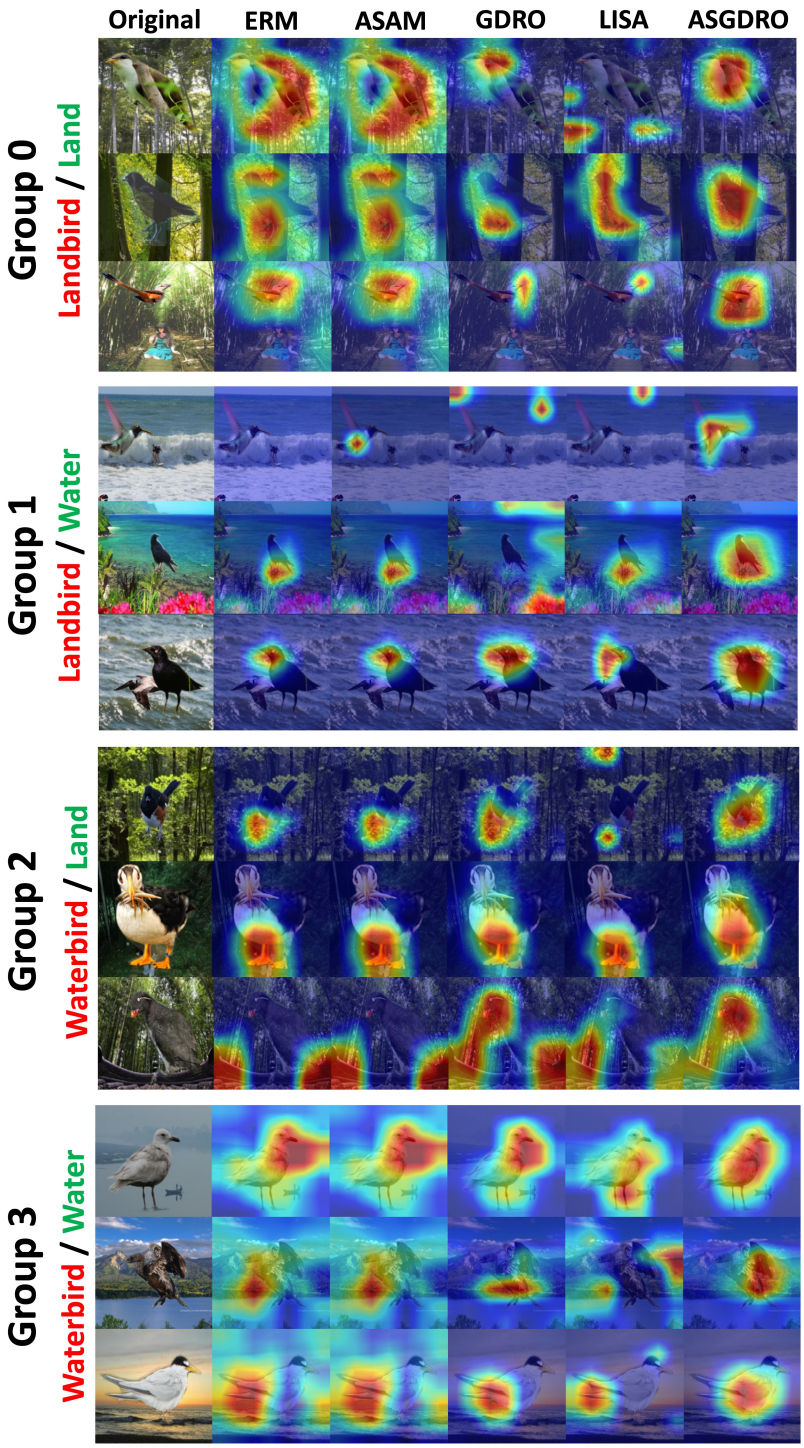

Figure 8: **Grad-CAM results on the Waterbirds Dataset.** The words highlighted in red represent invariant features: Landbird and Waterbird. On the contrary, the words highlighted in green represent spurious features: Land and Water background. In the Training Set, Group 1 and Group 2 are minority groups with significantly fewer data samples compared to other groups.

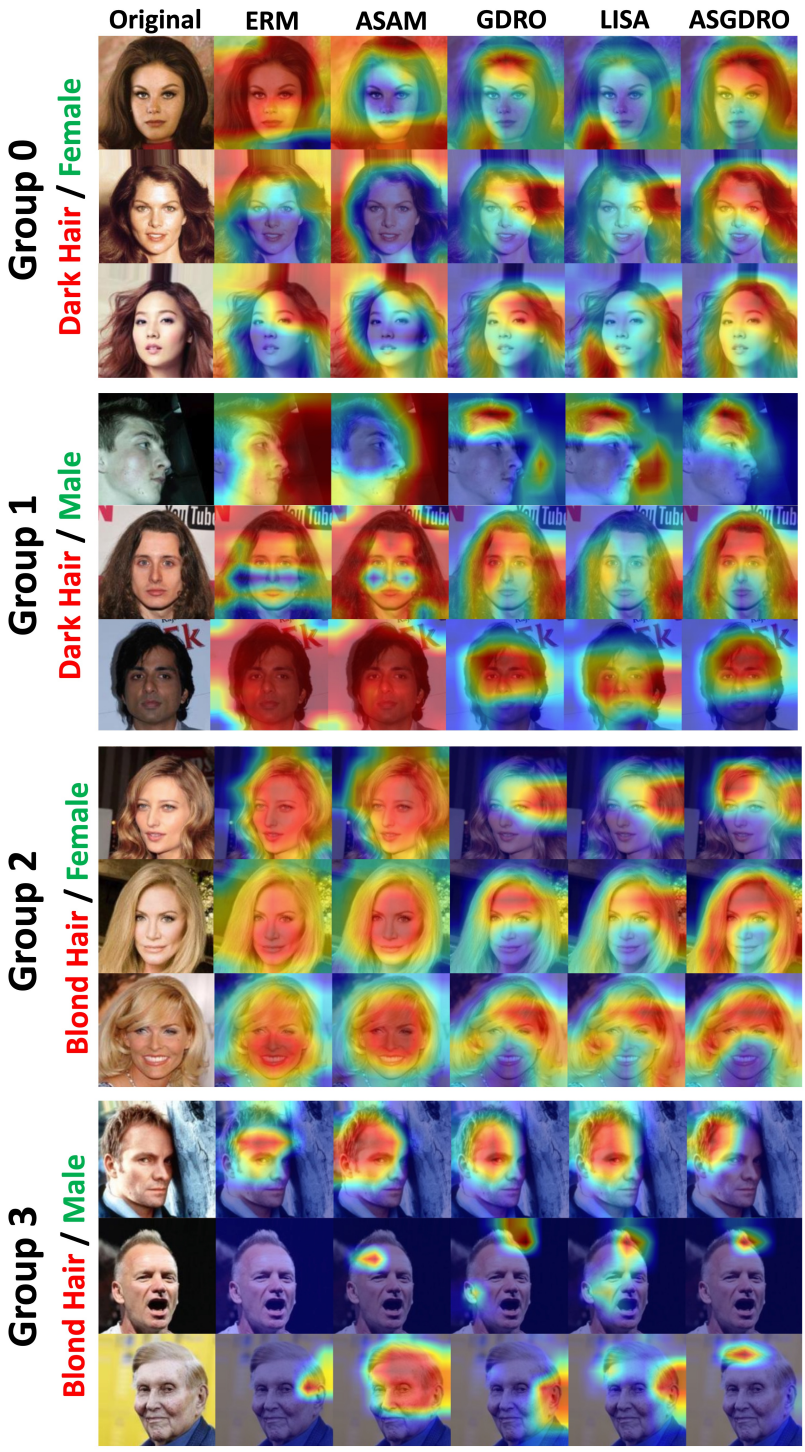

Figure 9: **Grad-CAM results on the CelebA Dataset.** The features highlighted in red represent invariant words: Dark Hair and Blond Hair. On the contrary, the words highlighted in green represent spurious features: Female and Male. In the Training Set, Group 3 is a minority group with significantly fewer data samples compared to other groups.

A.11 Hessian Analysis for Waterbirds Dataset

	The Largest Eigenvalue			The Second Largest Eigenvalue		
Method	Majority	Minority	Total	Majority	Minority	Total
ERM	990	4894	2265	166	511	709
ASAM	972	5475	2624	178	524	647
GDRO	131	447	353	118	346	129
ASGDRO	**107**	**342**	**279**	**98**	**274**	**105**

Table 6: **Hessian Analysis on Waterbirds.** ASGDRO finds the common flat minima for both majority and minority groups.

ERM and ASAM have significantly sharper minima for the minority group compared to GDRO and ASGDRO due to the spurious correlation, although ASAM is designed to find flat minima. Compared to GDRO and other baselines, ASGDRO achieves the lowest eigenvalue in the first and second maximum eigenvalues for every group.

References

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent YF Tan. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv preprint arXiv:2110.03141*, 2021.

Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent YF Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *arXiv preprint arXiv:2205.14083*, 2022.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. *arXiv preprint arXiv:2202.03599*, 2022.

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *Advances in Neural Information Processing Systems*, 36:47032–47051, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

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

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. *Advances in neural information processing systems*, 36, 2024.

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

Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Bingzhe Wu, Changqing Zhang, and Jianhua Yao. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *arXiv preprint arXiv:2209.08928*, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Xin Zhang, Yusuke Iwasawa, Yutaka Matsuo, and Shixiang Shane Gu. Amortized prompt: Lightweight fine-tuning for clip in domain generalization. *arXiv preprint arXiv:2111.12853*, 2021.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

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

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3769–3778, 2023.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.