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

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

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:

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

This simplifies to:

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

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:

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

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:

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

□

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

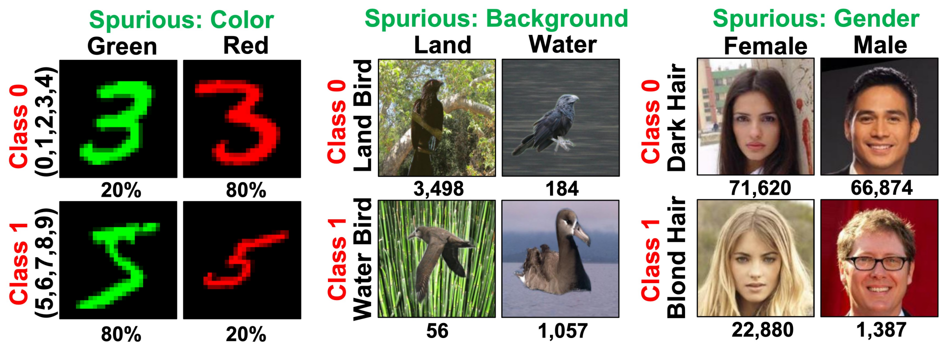

Figure 6: **CMNIST, Waterbirds, CelebA.** In each dataset, each row represents the class and each column represents the spurious feature. The numbers written below the images represent the ratio or count of data belonging to each group in the training dataset, where each group consists of (Class, Spurious 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 $\{10^{-6}, 2\times10^{-6}, 10^{-5}, 2\times10^{-5}\}$. 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

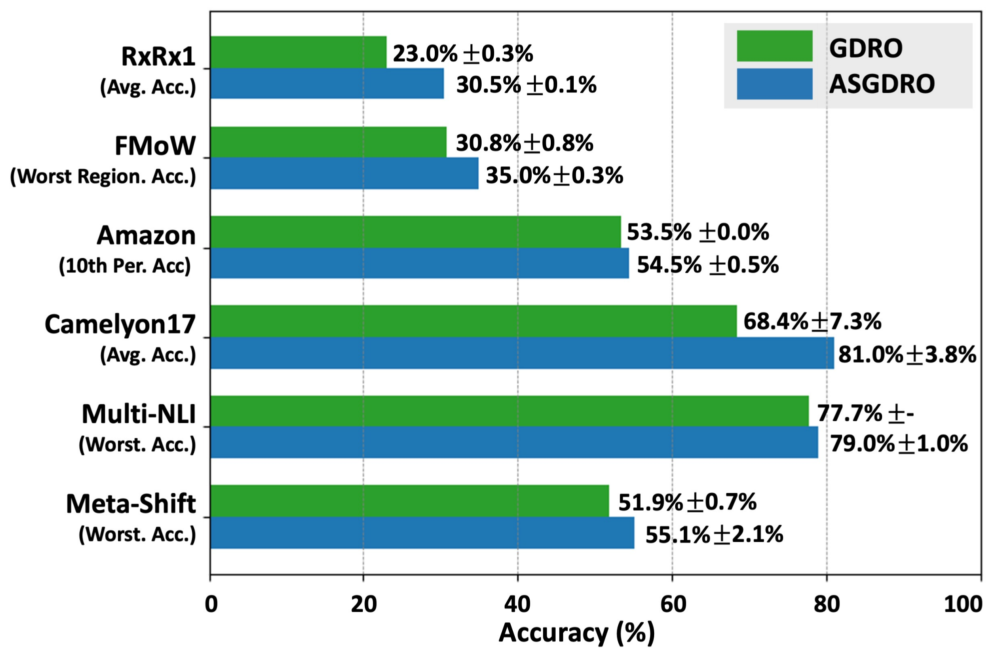

Figure 7: **Standard Deviations for Wilds Benchmark Datasets.**

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.

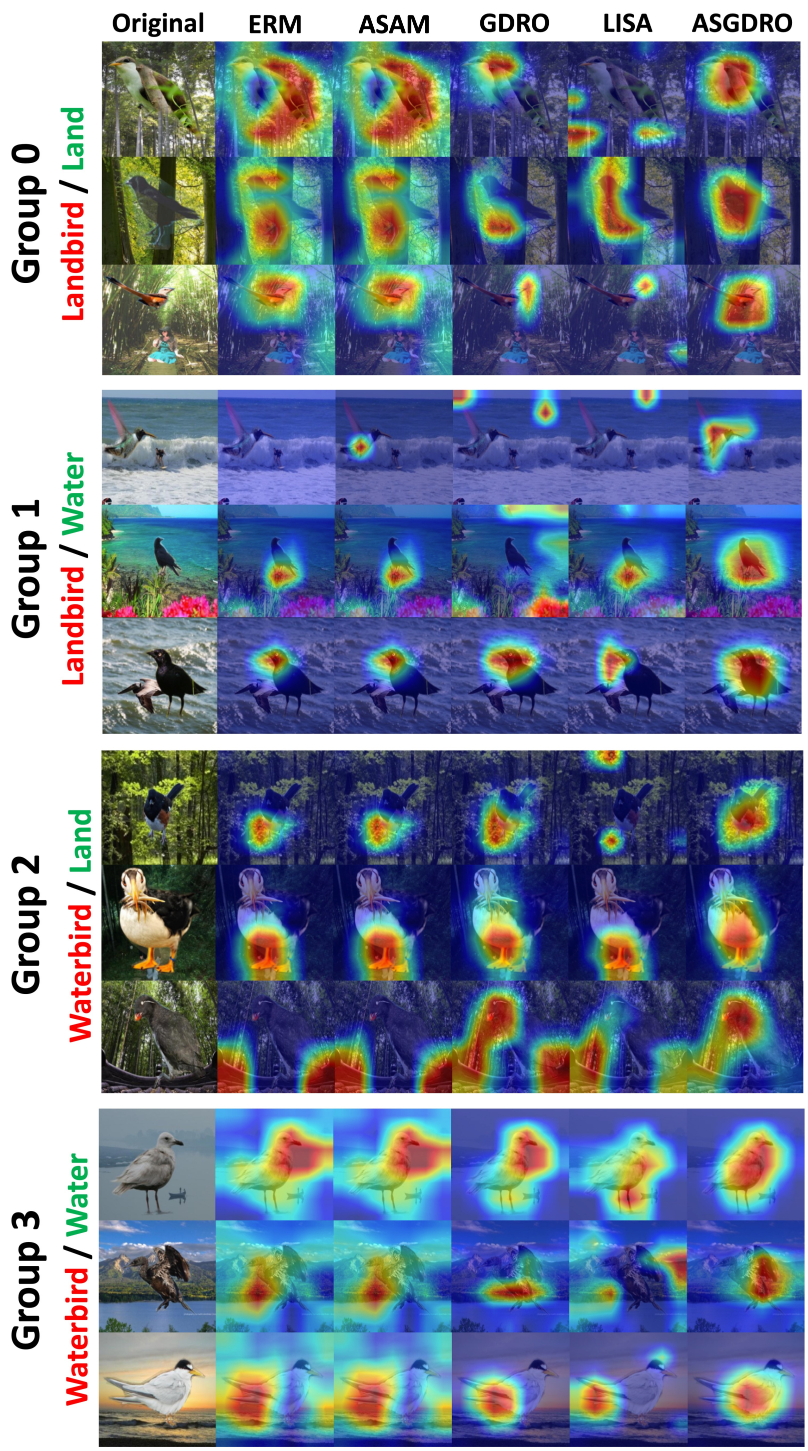

Figure 8: **Grad-CAM results on the Waterbirds Dataset.** The words highlighted in red represent invariant features: Landbird and Waterbird. On the contrary, the words highlighted in green represent spurious features: Land and Water background. In the Training Set, Group 1 and Group 2 are minority groups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

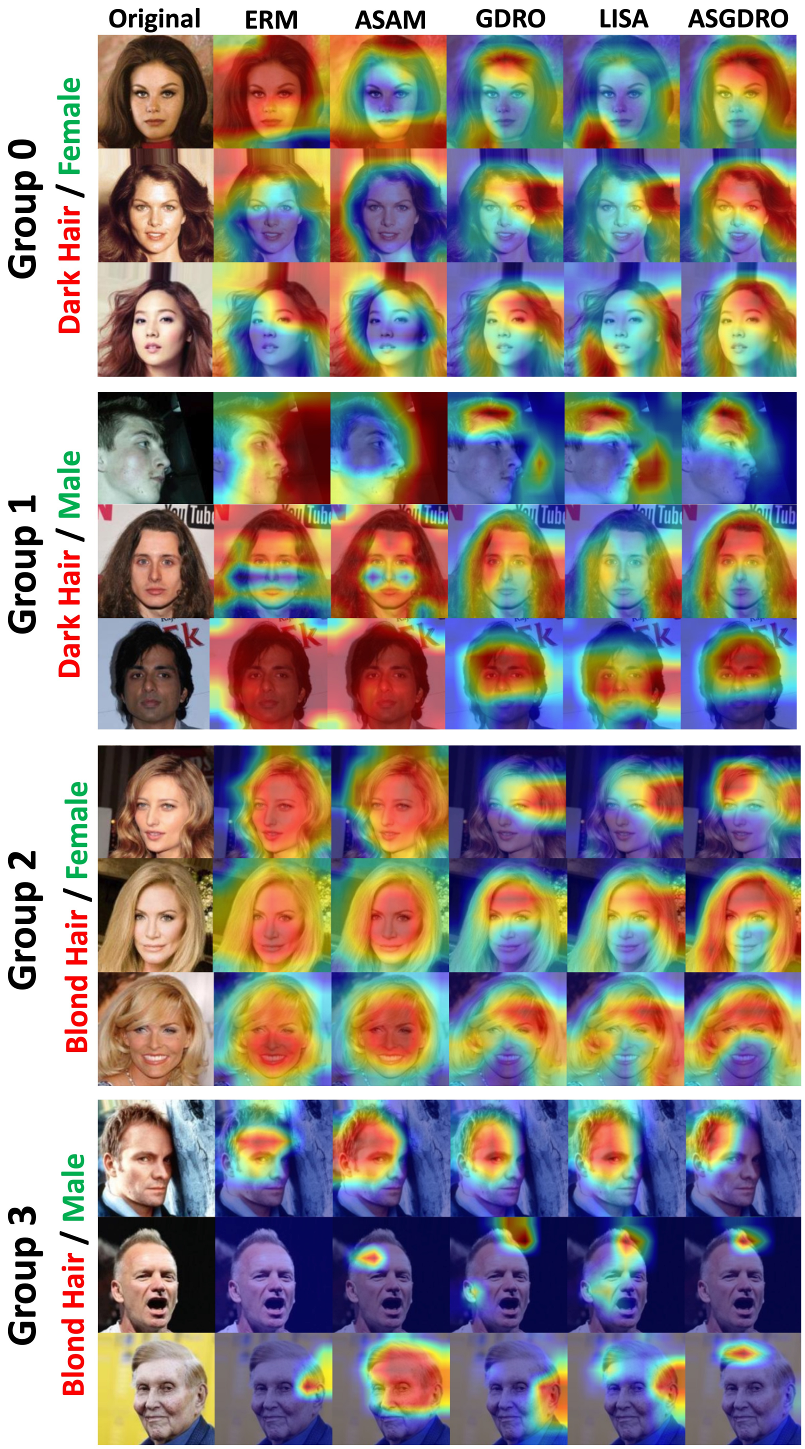

Figure 9: **Grad-CAM results on the CelebA Dataset.** The features highlighted in red represent invariant words: Dark Hair and Blond Hair. On the contrary, the words highlighted in green represent spurious features: Female and Male. In the Training Set, Group 3 is a minority group with significantly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