

# On the Connection between Invariant Learning and Adversarial Training for Out-of-Distribution Generalization

Shiji Xin<sup>1,4</sup>, Yifei Wang<sup>2</sup>, Jingtong Su<sup>3</sup>, Yisen Wang<sup>1,5\*</sup>

<sup>1</sup> Key Lab. of Machine Perception (MoE),

School of Intelligence Science and Technology, Peking University

<sup>2</sup> School of Mathematical Sciences, Peking University

<sup>3</sup> Center for Data Science, New York University

<sup>4</sup> School of EECS, Peking University

<sup>5</sup> Institute for Artificial Intelligence, Peking University

## Abstract

Despite impressive success in many tasks, deep learning models are shown to rely on spurious features, which will catastrophically fail when generalized to out-of-distribution (OOD) data. Invariant Risk Minimization (IRM) is proposed to alleviate this issue by extracting domain-invariant features for OOD generalization. Nevertheless, recent work shows that IRM is only effective for a certain type of distribution shift (*e.g.*, correlation shift) while it fails for other cases (*e.g.*, diversity shift). Meanwhile, another thread of method, Adversarial Training (AT), has shown better domain transfer performance, suggesting that it has the potential to be an effective candidate for extracting domain-invariant features. This paper investigates this possibility by exploring the similarity between the IRM and AT objectives. Inspired by this connection, we propose Domain-wise Adversarial Training (DAT), an AT-inspired method for alleviating distribution shift by domain-specific perturbations. Extensive experiments show that our proposed DAT can effectively remove domain-varying features and improve OOD generalization under both correlation shift and diversity shift.

## Introduction

Modern deep learning techniques have achieved remarkable success in many tasks (He et al. 2016; Wang et al. 2017; Brown et al. 2020). However, deep models will suffer catastrophic performance degradation under some scenarios, as they tend to exploit spurious correlations in the training data (Beery, Van Horn, and Perona 2018). One of those representative scenarios is the Out-of-Distribution (OOD) generalization, where the trained model is expected to perform well at the test time, even when the training and testing data come from different distributions (Zhang et al. 2021a). Another representative scenario in which deep models are unstable is the adversarial example. Researchers have found that deep models are quite brittle, as one can inject imperceptible perturbations into the input and cause the model to make wrong predictions with extremely high confidence (Szegedy et al. 2014).

These two issues have some similarities to each other. They both arise because deep networks do not learn the essential

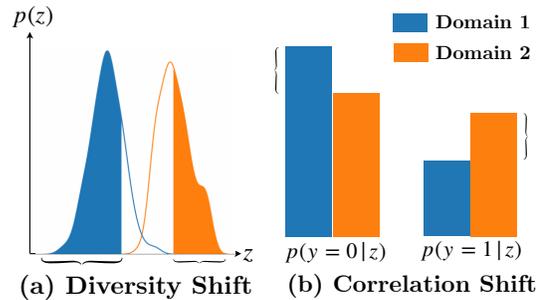


Figure 1: An illustrative example of the two kinds of distribution shifts. The curly brackets enclose the distribution shift between the environments. Here  $z$  stands for the spurious feature, and  $y$  stands for label class.

causal associations (or intrinsic features). Nevertheless, in their corresponding fields, different approaches have been proposed. A large class of methods called Invariant Causal Prediction (ICP) (Peters, Bühlmann, and Meinshausen 2016) is proposed for OOD generalization. Among them, Invariant Risk Minimization (IRM) (Arjovsky et al. 2019) attracts significant attention, which intends to extract invariant features across different data distributions and expects the model to ignore information related to the environment<sup>1</sup>. While for adversarial robustness against adversarial examples, Adversarial Training (AT) (Madry et al. 2018; Wang et al. 2019) is the most effective approach at the current stage (Athalye, Carlini, and Wagner 2018). It trains a model on adversarial examples generated by injecting perturbations optimized for each image into natural examples. Several recent works have explored the relationship between AT and OOD (Volpi et al. 2018; Shankar et al. 2018; Yi et al. 2021), but rarely focus on the typical domain generalization setting considered by IRM. Therefore, the two fields still seem rather independent. In this paper, we are going to explore their potential relationships.

Although IRM and its variants are promising on certain tasks, *e.g.*, CMNIST (Arjovsky et al. 2019), recent studies

\*Corresponding Author: Yisen Wang (yisen.wang@pku.edu.cn). Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>The terminologies of “domain”, “environment”, and “distribution” are often used interchangeably in current literature.

(Gulrajani and Lopez-Paz 2021) show that on a large-scale controlled experiment on OOD generalization, all these methods fail to exceed the simplest *i.i.d.* baseline, *i.e.*, Empirical Risk Minimization (ERM). Further, two kinds of distribution shifts in benchmark datasets are identified (Ye et al. 2022), *i.e.*, diversity shift and correlation shift, shown in Figure 1. Diversity shift refers to the shift of the distribution support of spurious feature  $z$ , for example, the style of the images changed from cartoon to sketch on object classification task. In contrast, correlation shift refers to the change in conditional probability (posterior distribution) of label  $y$  given spurious feature  $z$  on the same support, *e.g.*, the color in the CMNIST dataset (Ahuja et al. 2020). They found that an algorithm that performs well on one kind of distribution shift tends to perform poorly on the other one (Ye et al. 2022).

Thus, we need to seek better alternatives for OOD generalization, while AT seems to be a promising candidate from both theoretical and empirical aspects. Theoretically, by learning invariance *w.r.t.* local input perturbations (Wu, Xia, and Wang 2020), AT can be regarded as Distributionally Robust Optimization (DRO) (Sinha, Namkoong, and Duchi 2018; Volpi et al. 2018; Rahimian and Mehrotra 2019; Duchi, Glynn, and Namkoong 2021) over the  $\ell_p$ -bounded distributional shift. Thus, AT could reliably extract robust features, *e.g.*, the shape of the object, from the input (Ren et al. 2021). Empirically, several recent works show that AT has better domain transferability than ERM (Salman et al. 2020; Yi et al. 2021). These findings naturally lead to the following questions:

*Is AT related to IRM? If so, is AT helpful for OOD generalization?*

In this paper, we take a further step to answer these intriguing questions. We first reveal the connections between IRM and AT, and find that IRM can be regarded as an instance-reweighted version of Domain-wise Adversarial Training (DAT), a new version of adversarial training that we propose for multi-source domain generalization. Inspired by this connection, we further explore how DAT performs on OOD data. We first notice that DAT is suitable for solving domain generalization problems, as it can effectively remove relatively static background information with domain-wise perturbations. We further verify this intuition on both synthetic tasks and real-world datasets, where DAT shows clear advantages over ERM. Finally, we conduct extensive experiments on benchmark datasets and show that our DAT can consistently outperform ERM on tasks dominated by both correlation shift and diversity shift.

We summarize our contributions as follows:

- We theoretically derive the connection between IRM and AT. Based on this connection, we develop a new kind of adversarial training, Domain-wise Adversarial Training (DAT), for domain generalization.
- We analyze how DAT is beneficial for learning invariant features and verify our hypothesis through synthetic data and real-world datasets.
- Experiments on benchmark datasets show that DAT not only performs better than ERM under correlation shift

like IRM but also outperforms ERM under diversity shift like (sample-wise) AT.

## Related Works

**IRM and Its Variants.** Invariant Risk Minimization (IRM) (Arjovsky et al. 2019) develops a paradigm to extract causal (invariant) features and find the optimal invariant classifier on top of several given training environments. The work of Kamath et al. (2021) reveals the gap between IRM and IRMv1, showing that even in a simple model that echos the idea of the original IRM objective, IRMv1 can fail catastrophically. Rosenfeld, Ravikumar, and Risteski (2021) prove that when the number of training environments is not large enough, IRMv1 can face the risk of using environmental features.

**AT and Its Variants.** Szegedy et al. (2014) reported that one can inject imperceptible perturbations to fool deep models. Among the proposed defenses, Adversarial Training (AT) (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018; Wang et al. 2020; Wang and Wang 2022) is the promising and representative approach to training robust models (Athalye, Carlini, and Wagner 2018). Recently, Salman et al. (2020) showed that adversarially learned features could transfer better than standardly trained models, while various works (Volpi et al. 2018; Sinha, Namkoong, and Duchi 2018; Shankar et al. 2018; Ford et al. 2019; Qiao, Zhao, and Peng 2020; Yi et al. 2021; Gokhale et al. 2021) adopt sample-wise adversarial training or adversarial data augmentation to improve OOD robustness. However, most discussions are limited to distributional robustness *w.r.t.* Wasserstein distance. The small perturbations used in AT make it less practical to account for real-world OOD scenarios (*e.g.*, correlation to backgrounds), thus Wang et al. (2022) incorporated low-rank structured priors into AT for this kind of large data distribution shifts. Moreover, previous work shows that there seems to be a trade-off between the two distribution shifts: an algorithm that performs well on one task tends to perform poorly on the other (Ye et al. 2022). Instead, in this work, our proposed method can achieve fair performance on both correlation shift and diversity shift tasks.

## Relationship between IRM and AT Variants

**Notation.** Let  $\Phi : \mathcal{X} \subset R^n \rightarrow R^d$  denote the representation of a  $\theta$ -parameterized piecewise linear classifier, *i.e.*,  $\Phi(\cdot) = \phi^L(W^L \phi^{L-1}(\dots) + b^{L-1}) + b^L$ , where  $\phi^L$  is the activation function, and  $W^L, b^L$  denote the layer-wise weight matrix and bias vector, collectively denoted by  $\theta$ . Furthermore, let  $\beta$  be the linear classifier on the top, and let the network output be  $\beta \cdot \Phi(x) = \beta^\top \Phi(x)$ . Let  $\ell(\hat{y}, y) = -\log \sigma(y\hat{y})$  be the sample logistic loss. We consider a two-class ( $y = \pm 1$ ) classification setting with output dimension  $d = 1$ , and our discussion can be easily extended to general cases.

**ERM.** The traditional Empirical Risk Minimization (ERM) algorithm optimizes over the loss on *i.i.d.* data, *i.e.*,

$$\min_{\beta, \Phi} R(\beta \cdot \Phi), \text{ where } R(\beta \cdot \Phi) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell(\beta^\top \Phi(x), y). \quad (1)$$

In the OOD generalization problem, one faces a set of (training) *environments*  $\mathcal{E}$ , where each environment  $e \in \mathcal{E}$  corresponds to a unique data distribution  $D_e$ . When facing multiple environments, the ERM objective simply mixes the data together and takes the form

$$(\text{ERM}) \quad \min_{\beta, \Phi} \sum_e R^e(\beta \cdot \Phi), \quad (2)$$

where  $R^e(\beta \cdot \Phi) = \mathbb{E}_{(x,y) \sim D_e} \ell(\beta^\top \Phi(x), y)$ .

**IRM and IRMv1.** Instead of simply mixing the data together, IRM seeks to learn an *invariant* representation  $\Phi$  such that the objective can be minimized with the same classifier  $\beta$  in all training domains. Formally, we have

$$(\text{IRM}) \quad \begin{aligned} & \min_{\beta, \Phi} \sum_{e \in \mathcal{E}} R^e(\beta \cdot \Phi) \\ & \text{s.t. } \beta \in \arg \min_{\bar{\beta}} R^e(\bar{\beta} \cdot \Phi), \forall e \in \mathcal{E}. \end{aligned} \quad (3)$$

Since this bilevel optimization problem is difficult to solve, the practical version IRMv1 is formulated as regularized ERM, where the gradient penalty is calculated *w.r.t.* a dummy variable  $w$ :

$$(\text{IRMv1}) \quad \min_{\beta, \Phi} \sum_{e \in \mathcal{E}} \left[ R^e(\beta \cdot \Phi) + \underbrace{\lambda \|\nabla_w|_{w=1.0} R^e(w \cdot (\beta \cdot \Phi))\|^2}_{\text{Penalty}_{\text{IRM}}} \right]. \quad (4)$$

**AT.** Adversarial Training instead replaces ERM with a min-max objective,

$$\min_{\beta, \Phi} R^{\text{AT}}(\beta \cdot \Phi) = \min_{\beta, \Phi} \mathbb{E}_{(x,y) \sim D} \max_{\|\delta_x\|_p \leq \varepsilon} \ell(\beta^\top \Phi(x + \delta_x), y), \quad (5)$$

where one maximizes inner loss by injecting *sample-wise* perturbations  $\delta_x$  and solve the outer minimization *w.r.t.* parameters  $\beta, \Phi$  on the perturbed sample  $(x + \delta_x, y)$ . Typically, the perturbation is constrained within an  $\ell_p$  ball with radius  $\varepsilon$ . In this way, AT can learn models that are robust to  $\ell_p$  perturbations.

## Relating AT to IRM

As shown above, it seems that IRM and AT are two distinct learning paradigms, while, in fact, we can show that IRM is closely related to a certain kind of adversarial training. To see this, we first notice that AT can be rephrased into a regularized ERM loss with a penalty on *sample-wise* robustness through linearization:

$$\begin{aligned} R^{\text{AT}}(\beta \cdot \Phi) &= \mathbb{E}_{(x,y) \sim D} \max_{\|\delta_x\| \leq \varepsilon} \ell(\beta^\top \Phi(x + \delta_x), y) \\ &= \mathbb{E}_{(x,y) \sim D} \left[ \ell(\beta^\top \Phi(x), y) + \right. \\ & \quad \left. \max_{\|\delta_x\| \leq \varepsilon} (\ell(\beta^\top \Phi(x + \delta_x), y) - \ell(\beta^\top \Phi(x), y)) \right] \\ &\approx R(\beta \cdot \Phi) + \varepsilon \underbrace{\mathbb{E}_{(x,y) \sim D} \|\nabla_x \ell(\beta^\top \Phi(x), y)\|}_{\text{Penalty}_{\text{AT}}}, \end{aligned} \quad (6)$$

which resembles the gradient penalty adopted in IRMv1. One main difference is that AT's penalty is calculated *w.r.t.* sample-wise gradients, while IRM's penalty *w.r.t.* the population loss. This difference motivates us to adopt a population-level perturbation  $\delta$  instead.

---

### Algorithm 1: Domain-wise Adversarial Training

---

**Input:** Dataset of multiple environments  $D_e, e \in \mathcal{E}$ , desired  $\ell_p$  norm of the perturbation  $\varepsilon$ , perturbation step size  $\alpha$

**Output:** Model  $(\Phi, \beta)$

Randomly initiate  $\theta$ , perturbation  $\delta_e, \forall e \in \mathcal{E}$

**for** iterations = 1, 2, 3, ... **do**

**for** each environment  $e$  **do**

    1. Sample batch  $B_e$  from environment  $e$

    2. Update the perturbation  $\delta_e$  using one-step

      gradient ascent with step size  $\alpha$

    3. Project the perturbation  $\delta_e$  to a  $\ell_p$  ball of radius  $\varepsilon$

    4. Generate adversarial examples

$x_{adv} \leftarrow x + \delta_e, \forall x \in B_e$

    5. Update  $\Phi$  and  $\beta$  with gradient descent on  $x_{adv}$

**end for**

**end for**

---

**Proposed DAT.** Inspired by the above connection, we propose DAT, which adopts a *domain-wise* perturbation  $\delta_e$  for each training domain  $e \in \mathcal{E}$ . Formally, we have

$$\begin{aligned} & \min_{\beta, \Phi} \sum_{e \in \mathcal{E}} \mathbb{E}_{(x,y) \sim D_e} \ell(\beta^\top \Phi(x + \delta_e), y) \\ & \text{s.t. } \delta_e \in \arg \max_{\|\delta\| \leq \varepsilon} \mathbb{E}_{(x,y) \sim D_e} \ell(\beta^\top \Phi(x + \delta), y), \forall e \in \mathcal{E}. \end{aligned} \quad (7)$$

where the perturbation  $\delta_e$  is defined at the distribution level.

In practice, we solve the above problem by alternating updates of model parameters  $\beta, \Phi$  and perturbations  $\delta_e$ . Specifically, for each mini-batch  $B_e$  sampled from domain  $\mathcal{D}_e$ , we update  $\delta_e$  with  $B_e$  using gradient ascent to find the best adversarial perturbations. The adversarial samples are then used to train the model. A detailed description is in Algorithm 1.

In the setting of single-source domain generalization, where only one training domain is provided, DAT degenerates into universal adversarial training (UAT) (Moosavi-Dezfooli et al. 2017), where a single perturbation is provided for the entire training distribution.

## Connection between IRM and DAT

Here, we establish a formal connection between IRM and DAT. To begin with, we note that DAT can also be reformulated as a regularized ERM in the multi-domain scenario.

$$\begin{aligned} R^{\text{DAT}}(\beta \cdot \Phi) &= \sum_{e \in \mathcal{E}} \max_{\|\delta_e\| \leq \varepsilon} \mathbb{E}_{(x,y) \sim D_e} \ell(\beta^\top \Phi(x + \delta_e), y) \\ &= \sum_{e \in \mathcal{E}} \left[ \mathbb{E}_{(x,y) \sim D_e} \ell(\beta^\top \Phi(x), y) + \right. \\ & \quad \left. \max_{\|\delta_e\| \leq \varepsilon} \mathbb{E}_{(x,y) \sim D_e} (\ell(\beta^\top \Phi(x + \delta_e), y) - \ell(\beta^\top \Phi(x), y)) \right] \\ &\approx \sum_{e \in \mathcal{E}} \left[ R^e(\beta \cdot \Phi) + \varepsilon \underbrace{\mathbb{E}_{(x,y) \sim D_e} \|\nabla_x \ell(\beta^\top \Phi(x), y)\|}_{\text{Penalty}_{\text{DAT}}} \right]. \end{aligned} \quad (8)$$

Based on this reformulation, we can show that there exists an intrinsic relationship between DAT and IRM as in the following proposition:

**Proposition 0.1.** Consider each  $D_e$  as the corresponding distribution of a particular training domain  $e$ . For any  $\beta \cdot \Phi$  as a deep network with any activation function, the penalty term of IRMv1,  $\text{Penalty}_{\text{IRM}}$  (Eq. 4), could be expressed as the square of a reweighted version of the penalty term of the above approximate target,  $\text{Penalty}_{\text{DAT}}$  (Eq. 8), on each environment  $e$  with coefficients related to the distribution  $D_e$ , which could be stated as follows:

$$\text{Penalty}_{\text{IRM}} = \left\| \mathbb{E}_{D_e} [L_x x + \tilde{B}_x] \right\|^2 \quad (9)$$

$$\text{Penalty}_{\text{DAT}} = \left\| \mathbb{E}_{D_e} L_x \right\|^2 \quad (10)$$

where  $L_x = (1 - \sigma(y\beta^\top \Phi_x x))y\beta^\top \Phi_x$  and  $\tilde{B}_x = (1 - \sigma(y\beta^\top \Phi_x x))y\beta^\top B_x$ .  $B_x$  denotes the collection of constants introduced by bias terms in neural network layers.

If we consider the extreme case where each domain only contains one sample, we can see that DAT degenerates into AT as a special case. The equivalence between IRMv1 and (linearized) DAT in this setting can be shown as follows, which extends the similarity between IRM (Eq. 4) and AT (Eq. 6).

**Remark 0.2** (Equivalence under Single-sample Environments). When the environments degenerate into a single data point, we have the following relationship: If  $\varepsilon$  is sufficiently small, then for  $\beta \cdot \Phi$  as a deep network with any activation function, the penalty term of IRMv1 (Eq. 4) on each sample and the square of the maximization term of the linearized version of Eq. 7 (LDAT, obtained by the first-order approximation of DAT)

$$\text{Penalty}_{\text{LDAT}} = \left\langle \nabla_x \ell(\beta^\top \Phi(x), y), \pm \hat{\delta}_x \right\rangle \quad (11)$$

on each sample with perturbation  $\hat{\delta}_x = \pm \varepsilon x$  only differ by a fixed multiple  $\varepsilon^2$  and a bias term  $B_x$ , which is formally stated as

$$\begin{aligned} \text{Penalty}_{\text{LDAT}}^2 &= \left[ \left\langle \nabla_x \ell(\beta^\top \Phi(x), y), \pm \varepsilon x \right\rangle \right]^2 \\ &= \varepsilon^2 (1 - \sigma(\beta^\top \Phi(x)))^2 \left\| \beta^\top \Phi_x x \right\|^2 \quad (12) \\ &= \varepsilon^2 \cdot \text{Penalty}'_{\text{IRM}}, \end{aligned}$$

where  $\text{Penalty}'_{\text{IRM}} = (1 - \sigma(y\beta^\top \Phi(x)))^2 \left\| \beta^\top (\Phi_x x) \right\|^2$ ,  $\text{Penalty}_{\text{IRM}} = (1 - \sigma(y\beta^\top \Phi(x)))^2 \left\| \beta^\top (\Phi_x x + B_x) \right\|^2$ .

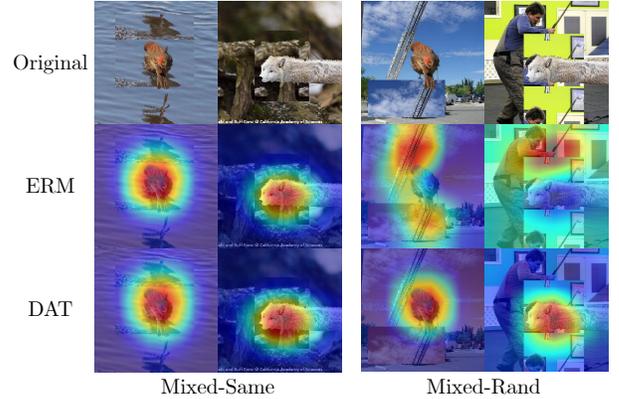
The proofs of Proposition 0.1 and Remark 0.2 can be found in Appendix. The above connection between DAT and IRM highlights that our DAT is potentially helpful in addressing OOD problems.

## Empirical Investigation on Domain-Wise Adversarial Training

In this section, we further explore how domain-wise perturbations could help alleviate distribution shifts in real-world



(a) Images and heatmaps of ERM from NICO.



(b) Images and heatmaps of ERM and DAT from Mixed-Same and Mixed-Rand.

Figure 2: Images from NICO and Mixed datasets and corresponding heatmaps of the models. The red regions in the heatmaps correspond to the model’s focus during prediction. Heatmaps demonstrate the effectiveness of DAT in attending to interested objects rather than uncorrelated background information compared to ERM.

scenarios. In particular, we noticed that domain-wise perturbations could effectively remove the domain-varying *background information*, which usually corresponds to spurious features for image classification tasks. We empirically verify this property by applying DAT to a well-designed OOD task based on background shift.

## Comparing ERM to DAT for Background Removal

**ERM Learns Spurious Background Features.** Our understanding of DAT is based on the idea that an image is composed of a foreground object and a corresponding background, and typically the object is the invariant feature, while the background is only spuriously correlated with the labels. However, models that rely on spurious background information will easily fail when encountering images from a different domain. This phenomenon is also empirically verified by Xiao et al. (2021), who find that models trained on an ImageNet-like dataset with ERM require image backgrounds to correctly classify large portions of test sets. These findings point out the limitations of ERM and motivate us to find a

solution that could effectively learn a background-invariant classifier.

**Removing Background Information with DAT.** Compared to the failure of ERM above, we notice that DAT can be applied to eliminate domain-wise background information with its domain-specific perturbations. Take the NICO images in Figure 2a as an example, where samples from the environment “on grass” have a common background dominated by green grass with low frequency, while the foreground object (e.g., the cats) has complex and instance-specific shape and texture with much higher frequency. In fact, Moosavi-Dezfooli et al. (2017) show that a universal perturbation vector lies in a low-dimensional subspace, which fits the background statistics and could be used to eliminate the low-frequency background factor. Therefore, when our DAT is applied to these samples, the domain-wise perturbation will capture the common domain-specific background. Consequently, domain-wise AT will help remove the dependence on these spurious background features.

### Empirical Verification with Controlled Experiments

We construct a synthetic OOD task to verify the above analysis by evaluating a classifier’s dependence on background information. It is based on two datasets introduced by Xiao et al. (2021), Mixed-Same and Mixed-Rand, which are constructed from a subset of ImageNet images with the background of each image replaced by another background that is either of the same class (Mixed-Same) or a random class (Mixed-Rand). As they are perfect candidates for evaluating a classifier in terms of its background dependence, we construct a new OOD task by using Mixed-Same as the training domain and evaluating the learned classifier on Mixed-Rand as the test domain. If the classifier relies heavily on background information, it will perform poorly in the test domain, where objects and backgrounds are disentangled. In particular, the experiment results show that ERM achieves a test accuracy of 71.9%, while DAT achieves 72.6% in the test domain with random background, which means that DAT has a better generalization ability by effectively removing background information. Sample images from the dataset and the corresponding attention heatmaps are shown in Figure 2b, demonstrating that ERM may lose focus when the background correlation is broken while DAT does not. Details of the experiment are shown in Appendix.

### Experiments

For the experiments, we follow the setting in Ye et al. (2022) and evaluate the OOD generalization on both types of distribution shift: diversity shift and correlation shift. In particular, we select four representative tasks. For the correlation shift, we use CMNIST (Arjovsky et al. 2019), a synthetic dataset on digit classification with color as the spurious feature, and NICO (He, Shen, and Cui 2020), a real-world dataset on object classification with context as the spurious feature. Regarding diversity shift, we use PACS (Li et al. 2017) and Terra Incognita (Beery, Van Horn, and Perona 2018), which are both datasets consisting of natural images with four domains. To ensure fair evaluation, we perform all of our experiments following the evaluation pipeline of DomainBed (Gulrajani

and Lopez-Paz 2021). Specifically, we use the same dataset splitting and model selection strategy as in Ye et al. (2022) for each task. For datasets except for NICO, one of the domains is used as the test domain. We train the models in each run, treating one of the domains as the test domain and the rest as training domains, then report the average accuracy of all domains. For NICO, the training, test, and evaluation domains are predefined. We train the models on training domains, evaluate them on the evaluation domain for model selection, and report their accuracy on the test domain. More details of the experimental settings and domain-split results can be found in Appendix.

When training models using DAT, we first perform standard data augmentation (Gulrajani and Lopez-Paz 2021), then proceed with the update of the perturbation and model parameters as shown in Algorithm 1, where the perturbed samples are clipped to the legal range after data augmentation. We use DAT with perturbation bounded by  $\ell_2$ -norm in our experiments. To test DAT on a wider range of tasks, we also carry out experiments on DomainBed (Gulrajani and Lopez-Paz 2021), the results are shown in Appendix.

### Evaluation on Benchmark Datasets

We compare our results with previous work, including vanilla ERM, invariance-based methods including IRM (Ahuja et al. 2020), robust optimization methods including GroupDRO (Sagawa et al. 2020), distribution matching methods including MMD (Li et al. 2018b) and CORAL (Sun and Saenko 2016), a method based on domain classifier DANN (Ganin et al. 2016), and various other algorithms. The results of the CMNIST dataset are adopted from Gulrajani and Lopez-Paz (2021), which is the average of three domains, while the results of the other datasets are adopted from Ye et al. (2022). In addition to that, we implement and test four AT-based algorithms, including sample-wise adversarial training AT (Goodfellow, Shlens, and Szegedy 2015), universal adversarial training UAT (Shafahi et al. 2020), WRM (Sinha, Namkoong, and Duchi 2018), and adversarial data augmentation ADA (Volpi et al. 2018).

From Table 1, we can see that DAT consistently outperforms ERM and achieves good performance under both diversity and correlation shifts. Specifically, DAT achieves much better results than ERM on correlation-shift tasks (CMNIST and NICO) like IRM. Second, we can see that most domain generalization algorithms at the moment cannot surpass ERM on tasks dominated by diversity shifts. Although RSC has great performance on these tasks, it performs much worse under correlation shifts. However, DAT consistently outperforms ERM and could account for both kinds of shifts. Furthermore, compared to other AT-based algorithms (i.e., sample-wise AT, UAT, WRM, and ADA), DAT has a fair performance by considering a domain-wise perturbation that removes domain-varying spurious features. The results demonstrate the effectiveness of DAT in dealing with domain discrepancy.

### Extension to Single Domain Generalization

As discussed in Section , DAT can reduce the influence of background even when no domain labels are given, which

Algorithm	Correlation shift		Diversity shift	
	CMNIST	NICO	PACS	TerraInc
ERM (Baseline)	58.5 ± 0.3	72.1 ± 1.6	81.5 ± 0.0	42.6 ± 0.9
VREx (Krueger et al. 2021)	56.3 ± 1.9	71.5 ± 2.3	81.8 ± 0.1	40.7 ± 0.7
GroupDRO (Sagawa et al. 2020)	61.2 ± 0.6	71.0 ± 0.4	80.4 ± 0.3	36.8 ± 1.1
IRM (Ahuja et al. 2020)	<b>70.2 ± 0.2</b>	<b>73.3 ± 2.1</b>	81.1 ± 0.3	42.0 ± 1.8
ARM (Zhang et al. 2021b)	63.2 ± 0.7	67.3 ± 0.2	81.0 ± 0.4	39.4 ± 0.7
RSC (Huang et al. 2020)	58.5 ± 0.5	<b>74.3 ± 1.9</b>	<b>82.8 ± 0.4</b>	<b>43.6 ± 0.5</b>
DANN (Ganin et al. 2016)	58.3 ± 0.2	69.4 ± 1.7	81.1 ± 0.4	39.5 ± 0.2
MMD (Li et al. 2018b)	63.4 ± 0.7	68.9 ± 1.2	81.7 ± 0.2	38.3 ± 0.4
MTL (Blanchard et al. 2021)	57.6 ± 0.3	70.6 ± 0.8	81.2 ± 0.4	38.9 ± 0.6
MLDG (Li et al. 2018a)	58.4 ± 0.2	66.6 ± 2.4	73.0 ± 0.4	27.3 ± 2.0
SagNet (Nam et al. 2021)	58.2 ± 0.3	69.8 ± 0.7	81.6 ± 0.4	42.3 ± 0.7
CORAL (Sun and Saenko 2016)	57.6 ± 0.5	70.8 ± 1.0	81.6 ± 0.6	38.3 ± 0.7
Mixup (Yan et al. 2020)	58.4 ± 0.2	72.5 ± 1.5	79.8 ± 0.6	39.8 ± 0.3
AT (sample-wise) (Goodfellow, Shlens, and Szegedy 2015)	57.9 ± 0.4	70.5 ± 0.7	<b>82.0 ± 0.2</b>	42.6 ± 0.3
UAT (Shafahi et al. 2020)	58.7 ± 2.3	69.1 ± 1.2	80.7 ± 0.4	41.9 ± 1.8
WRM (Sinha, Namkoong, and Duchi 2018)	57.9 ± 3.3	68.2 ± 1.0	80.4 ± 0.0	26.1 ± 1.5
ADA (Volpi et al. 2018)	56.3 ± 0.4	69.5 ± 1.9	80.2 ± 0.2	41.2 ± 0.7
DAT (our work)	<b>68.4 ± 2.0</b>	72.6 ± 1.7	<b>82.0 ± 0.1</b>	<b>42.7 ± 0.7</b>

Table 1: Test accuracy (%) on four representative tasks for OOD generalization. According to the OOD-Bench (Ye et al. 2022), two are dominated by correlation shift, CMNIST and NICO, and two are dominated by diversity shift, PACS and TerraInc. We highlight the top two results on each task.

corresponds to the single-source domain generalization setting. We conduct experiments that strengthen this claim to see how DAT could help in the single-source domain generalization setting. The experimental setting can be found in Appendix.

The results on the four datasets used in Table 1 are shown in Table 2. We can see that our DAT has fair performance on all four datasets compared to ERM, either under correlation shift or diversity shift. Although the difference is not as significant as in the multiple-domain setting, it shows that our DAT works for both single-domain and multi-domain generalization scenarios. In particular, its advantages are more significant with multiple domains, where the domain-wise perturbation mechanism is more effective.

We also test DAT on Digits, a common single-source domain generalization dataset consisting of five sub-datasets: MNIST (LeCun et al. 1998), MNIST-M (Ganin and Lempitsky 2015), SVHN (Netzer et al. 2011), SYN (Ganin and Lempitsky 2015), and USPS (Denker et al. 1988). We show the results in Table 3. The performance of ERM is adopted from Qiao, Zhao, and Peng (2020). From these results, we can see that DAT retains its better generalization ability in the challenging single-domain generalization setting by outperforming ERM in all tasks and could be a promising alternative to ERM.

Algorithm	CMNIST	NICO	PACS	TerraInc
ERM	45.8 ±0.7	63.2 ±2.9	59.8 ±1.6	27.3 ±4.0
DAT	<b>46.0</b> <b>±2.0</b>	<b>64.4</b> <b>±0.7</b>	<b>59.9</b> <b>±2.4</b>	<b>28.1</b> <b>±4.0</b>

Table 2: Average test accuracy (%) of ERM, DAT on four representative tasks (single-source domain generalization).

	SVHN	MNIST-M	SYN	USPS	Avg
ERM	27.83	52.72	39.65	76.94	49.29
DAT	<b>28.2</b> <b>±1.0</b>	<b>55.8</b> <b>±1.4</b>	<b>43.1</b> <b>±0.2</b>	<b>81.3</b> <b>±1.0</b>	<b>52.1</b>

Table 3: Results of ERM and DAT on Digits dataset

Algorithm	CMNIST	NICO
UAT	58.7 ± 2.3	69.1 ± 1.2
Ensemble UAT	58.2 ± 2.3	60.8 ± 0.2
DAT	<b>68.4 ± 2.0</b>	<b>72.6 ± 1.7</b>

Table 4: Results of UAT, Ensemble UAT, and DAT on CMNIST and NICO.

Under this setting, DAT generates perturbations in the only training domain, thus degenerating into UAT. However, as shown in Table 1, DAT performs much better in the multidomain setting, which shows the necessity of generating

domain-specific perturbations. Furthermore, to show that models trained using DAT on multi-domains are not a trivial ensemble of models trained on single-domains, we train voting classifiers on CMNIST and NICO using UAT models trained separately on each domain. The results are shown in Table 4. We can see that neither UAT nor ensemble UAT has the same performance as DAT. This verifies the effectiveness of learning domain-wise perturbations using DAT.

### Analysis

We conduct extensive experiments to better understand what our algorithm learns and how the magnitude of hyperparameters affects its performance.

**Qualitative Analysis through Semantic Graphs.** We use GradCam (Selvaraju et al. 2017; Gildenblat and contributors 2021) to visualize the attention heatmaps of models trained by ERM, sample-wise AT, and DAT on the NICO dataset. The results are shown in Figure 3, where we can see that DAT pays more attention to the object itself than the strongly correlated background, while ERM and sample-wise AT tend to use environmental features instead.

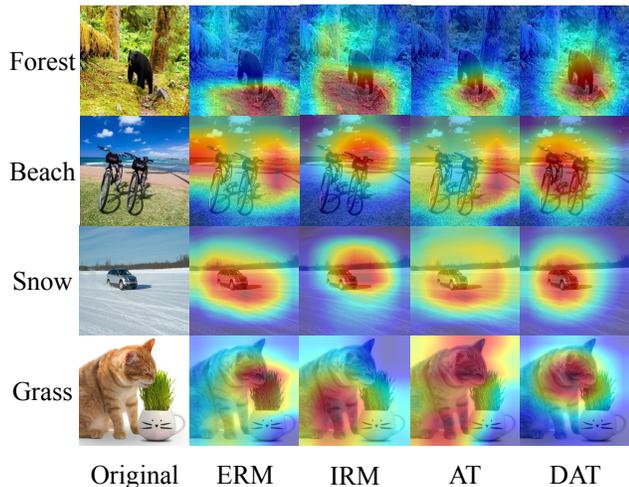


Figure 3: Attention heatmaps of ERM, IRM, (sample-wise) AT, and our DAT on the NICO dataset. Compared to other methods, DAT focuses more precisely on the object itself.

**Perturbation Radius and Step Size.** We investigate the effect of the perturbation radius  $\varepsilon$  and the perturbation step size  $\alpha$  on the NICO dataset. The results are shown in Table 5. The results show that the perturbation radius  $\varepsilon$  greatly affects the OOD performance. When  $\varepsilon$  is too large ( $> 10^{-1}$ ), it begins to hurt the invariant feature and causes performance degradation (from 72.9% to 67.9%). The step size  $\alpha$  has a smaller influence, and choosing a value between 1/100 and 1/10 times the size of  $\varepsilon$  would be appropriate. The effect of the norm used for perturbation in DAT is analyzed in Appendix.

### Discussions

**Comparison with Sample-wise AT.** Previous works (Hendrycks et al. 2021; Yi et al. 2021) try to exploit *sample-wise* AT as a data augmentation strategy to obtain higher

Radius $\varepsilon$	Step Size $\alpha$	Acc (%)
$[10^{-2}, 10^{-1}]$	$[10^{-4}, 10^{-3}]$	$72.6 \pm 1.7$
	$[10^{-3}, 10^{-2}]$	$72.0 \pm 2.1$
	$[10^{-2}, 10^{-1}]$	$68.9 \pm 1.5$
$[10^{-1}, 10^0]$	$[10^{-3}, 10^{-2}]$	$71.2 \pm 0.4$
	$[10^{-2}, 10^{-1}]$	$66.6 \pm 1.7$
	$[10^{-1}, 10^0]$	$69.4 \pm 0.4$
$[10^0, 10^1]$	$[10^{-3}, 10^{-2}]$	$64.6 \pm 0.3$
	$[10^{-2}, 10^{-1}]$	$67.9 \pm 2.0$
	$[10^{-1}, 10^0]$	$67.4 \pm 0.4$

Table 5: Comparison of the test accuracy of different perturbation radius  $\varepsilon$  and step size  $\alpha$  of  $\ell_2$ -norm bounded DAT on the NICO dataset.

OOD performance. However, the performance only improves when the distribution shift is dominated by diversity shift, *e.g.*, noise, and blurring. Otherwise, performance might be degraded, as shown in Table 1. One possible explanation is that sample-wise AT fails to capture the domain-level variations as DAT. As a result, it may add perturbations to the invariant features and hurt performance, especially under correlation shift.

**Comparison with Invariant Causal Prediction.** A thread of methods, including ICP (Peters, Bühlmann, and Meinshausen 2016), IRM (Arjovsky et al. 2019), and IGA (Koyama and Yamaguchi 2020), try to find invariant data representations that could induce an invariant classifier. They have superior performance on synthetic datasets like CMNIST but fail to outperform ERM on real-world datasets (tasks dominated by both correlation shift and diversity shift). We believe that these failures could be attributed to the lack of prior information of their invariant learning principles. In our DAT, we effectively exploit the foreground-background difference in image classification tasks through domain-wise perturbations.

### Conclusion

In this work, we carefully analyze the similarity between IRM and adversarial training in a domain-wise manner and establish a formal connection between OOD and adversarial robustness. Based on this connection, we propose a new adversarial training method for domain generalization: Domain-wise Adversarial Training (DAT). We show that it could effectively remove spurious background features in image classification and obtain fair performance on benchmark datasets. In particular, our DAT could consistently outperform ERM on tasks dominated by both the correlation shift and the diversity shift, while previous methods typically fail in one of the two cases.

### Acknowledgments

Yisen Wang is partially supported by the NSF China (No. 62006153), Open Research Projects of Zhejiang Lab (No. 2022RC0AB05), and Huawei Technologies Inc.

## References

- Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020. Invariant risk minimization games. In *ICML*.
- Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. arXiv:1907.02893.
- Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*.
- Beery, S.; Van Horn, G.; and Perona, P. 2018. Recognition in terra incognita. In *ECCV*.
- Blanchard, G.; Deshmukh, A. A.; Dogan, Ü.; Lee, G.; and Scott, C. D. 2021. Domain Generalization by Marginal Transfer Learning. *J. Mach. Learn. Res.*, 22: 2:1–2:55.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.
- Denker, J. S.; Gardner, W. R.; Graf, H. P.; Henderson, D.; Howard, R. E.; Hubbard, W. E.; Jackel, L. D.; Baird, H. S.; and Guyon, I. 1988. Neural Network Recognizer for Hand-Written Zip Code Digits. In *NeurIPS*.
- Duchi, J. C.; Glynn, P. W.; and Namkoong, H. 2021. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*.
- Ford, N.; Gilmer, J.; Carlini, N.; and Cubuk, E. D. 2019. Adversarial Examples Are a Natural Consequence of Test Error in Noise. In *ICML*.
- Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.
- Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17: 59:1–59:35.
- Gildenblat, J.; and contributors. 2021. PyTorch library for CAM methods. <https://github.com/jacobgil/pytorch-gradcam>. Accessed: 2023-03-10.
- Gokhale, T.; Anirudh, R.; Kailkhura, B.; J. Thiagarajan, J.; Baral, C.; and Yang, Y. 2021. Attribute-Guided Adversarial Training for Robustness to Natural Perturbations. In *AAAI*.
- Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*.
- Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *ICLR*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- He, Y.; Shen, Z.; and Cui, P. 2020. Towards Non-IID Image Classification: A Dataset and Baselines. *Pattern Recognition*, 107383.
- Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *ICCV*.
- Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-Challenging Improves Cross-Domain Generalization. In *ECCV*.
- Kamath, P.; Tangella, A.; Sutherland, D. J.; and Srebro, N. 2021. Does Invariant Risk Minimization Capture Invariance? In *AISTATS*.
- Koyama, M.; and Yamaguchi, S. 2020. When is invariance useful in an Out-of-Distribution Generalization problem? arXiv:2008.01883.
- Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Priol, R. L.; and Courville, A. C. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REX). In *ICML*.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.
- Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *ICCV*.
- Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018a. Learning to generalize: Meta-learning for domain generalization. In *AAAI*.
- Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *CVPR*.
- Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.
- Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *CVPR*.
- Nam, H.; Lee, H.; Park, J.; Yoon, W.; and Yoo, D. 2021. Reducing Domain Gap by Reducing Style Bias. In *CVPR*.
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*.
- Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*.
- Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to Learn Single Domain Generalization. In *CVPR*.
- Rahimian, H.; and Mehrotra, S. 2019. Distributionally robust optimization: A review. arXiv:1908.05659.
- Ren, J.; Zhang, D.; Wang, Y.; Chen, L.; Zhou, Z.; Chen, Y.; Cheng, X.; Wang, X.; Zhou, M.; Shi, J.; et al. 2021. A unified game-theoretic interpretation of adversarial robustness. In *NeurIPS*.
- Rosenfeld, E.; Ravikumar, P. K.; and Risteski, A. 2021. The Risks of Invariant Risk Minimization. In *ICLR*.
- Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally Robust Neural Networks. In *ICLR*.
- Salman, H.; Ilyas, A.; Engstrom, L.; Kapoor, A.; and Madry, A. 2020. Do Adversarially Robust ImageNet Models Transfer Better? In *NeurIPS*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J. P.; Davis, L. S.; and Goldstein, T. 2020. Universal Adversarial Training. In *AAAI*.

Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; Jyothi, P.; and Sarawagi, S. 2018. Generalizing Across Domains via Cross-Gradient Training. In *ICLR*.

Sinha, A.; Namkoong, H.; and Duchi, J. 2018. Certifying Some Distributional Robustness with Principled Adversarial Training. In *ICLR*.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to Unseen Domains via Adversarial Data Augmentation. In *NeurIPS*.

Wang, H.; and Wang, Y. 2022. Self-Ensemble Adversarial Training for Improved Robustness. In *ICLR*.

Wang, Q.; Wang, Y.; Zhu, H.; and Wang, Y. 2022. Improving Out-of-Distribution Generalization by Adversarial Training with Structured Priors. In *NeurIPS*.

Wang, Y.; Deng, X.; Pu, S.; and Huang, Z. 2017. Residual convolutional CTC networks for automatic speech recognition. *arXiv preprint arXiv:1702.07793*.

Wang, Y.; Ma, X.; Bailey, J.; Yi, J.; Zhou, B.; and Gu, Q. 2019. On the Convergence and Robustness of Adversarial Training. In *ICML*.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In *ICLR*.

Wu, D.; Xia, S.-T.; and Wang, Y. 2020. Adversarial Weight Perturbation Helps Robust Generalization. In *NeurIPS*.

Xiao, K. Y.; Engstrom, L.; Ilyas, A.; and Madry, A. 2021. Noise or Signal: The Role of Image Backgrounds in Object Recognition. In *ICLR*.

Yan, S.; Song, H.; Li, N.; Zou, L.; and Ren, L. 2020. Improve unsupervised domain adaptation with mixup training. *arXiv:2001.00677*.

Ye, N.; Li, K.; Bai, H.; Yu, R.; Hong, L.; Zhou, F.; Li, Z.; and Zhu, J. 2022. OoD-Bench: Quantifying and Understanding Two Dimensions of Out-of-Distribution Generalization. In *CVPR*.

Yi, M.; Hou, L.; Sun, J.; Shang, L.; Jiang, X.; Liu, Q.; and Ma, Z.-M. 2021. Improved OOD Generalization via Adversarial Training and Pre-training. In *ICML*.

Zhang, D.; Ahuja, K.; Xu, Y.; Wang, Y.; and Courville, A. 2021a. Can subnetwork structure be the key to out-of-distribution generalization? In *ICML*.

Zhang, M.; Marklund, H.; Dhawan, N.; Gupta, A.; Levine, S.; and Finn, C. 2021b. Adaptive Risk Minimization: Learning to Adapt to Domain Shift. In *NeurIPS*.