

Invariant Information Bottleneck for Domain Generalization

Bo Li¹ Yifei Shen² Yezhen Wang¹
Wenzhen Zhu³ Colorado Reed⁴ Kurt Keutzer⁴
Dongsheng Li¹ Han Zhao⁵

¹ Microsoft Research Asia, China ² HKUST, China ³ WUSTL, USA
⁴ UC Berkeley, USA ⁵ UIUC, USA

Abstract

Invariant risk minimization (IRM) has recently emerged as a promising alternative for domain generalization. Nevertheless, the loss function is difficult to optimize for nonlinear classifiers and the original optimization objective could fail when pseudo-invariant features and geometric skews exist. Inspired by IRM, in this paper we propose a novel formulation for domain generalization, dubbed invariant information bottleneck (IIB). IIB aims at minimizing invariant risks for nonlinear classifiers and simultaneously mitigating the impact of pseudo-invariant features and geometric skews. Specifically, we first present a novel formulation for invariant causal prediction via mutual information. Then we adopt the variational formulation of the mutual information to develop a tractable loss function for nonlinear classifiers. To overcome the failure modes of IRM, we propose to minimize the mutual information between the inputs and the corresponding representations. IIB significantly outperforms IRM on synthetic datasets, where the pseudo-invariant features and geometric skews occur, showing the effectiveness of proposed formulation in overcoming failure modes of IRM. Furthermore, experiments on DomainBed show that IIB outperforms 13 baselines by 0.9% on average across 7 real datasets.

Introduction

In most statistical machine learning algorithms, a fundamental assumption is that the training data and test data are *independently and identically distributed* (i.i.d.). However, the data we have in many real-world applications are not i.i.d. Distributional shifts are ubiquitous. Under such circumstances, classic statistical learning paradigms with strong generalization guarantees, e.g., Empirical Risk Minimization (ERM) (Vapnik 1999), often fail to generalize due to the violation of the i.i.d. assumption. It has been widely observed that the performance of a model often deteriorates dramatically when it is faced with samples from a different domain, even under a mild distributional shift (Arjovsky et al. 2019). On the other hand, collecting training samples from all possible future scenarios is essentially infeasible. Hence, understanding and improving the generalization of models on *out-of-distribution* data is crucial.

Domain generalization (DG), which aims to learn a model from several different domains so that it generalizes to *un-*

seen related domains, has recently received much attention. From the perspective of representation learning, there are several paradigms towards this goal, including invariant representation learning (Muandet, Balduzzi, and Schölkopf 2013; Zhao et al. 2018; Tachet des Combes et al. 2020), invariant causality prediction (Arjovsky et al. 2019; Krueger et al. 2020b), meta-learning (Balaji, Sankaranarayanan, and Chellappa 2018; Du et al. 2020), and feature disentanglement (Du et al. 2020; Peng et al. 2019). Of particular interest is the invariant learning methods. Some early works, e.g., DANN (Ganin et al. 2017), CDANN (Long et al. 2018), aim at finding representations that are invariant across domains. Nevertheless, learning invariant representations fails for domain adaptation or generalization when the marginal label distributions change between source and target domains (Zhao et al. 2019a). Recently, Invariant Causal Prediction (ICP), and its follow-up Invariant Risk Minimization (IRM), have attracted much interest. ICP assumes that the data are generated according to a structural causal model (SCM) (Pearl 2010). The causal mechanism for the data generating process is the same across domains, while the *interventions* can vary among different domains. Under such data generative assumptions, IRM (Arjovsky et al. 2019) attempts to learn an optimal classifier that is invariant across domains. ICP then argues that under the SCM assumption, such a classifier can generalize across domains.

Despite the intuitive motivations, IRM falls short in several aspects. First, the proposed loss function in (Arjovsky et al. 2019) is difficult to optimize when the classifier is nonlinear. Furthermore, it has been shown that IRM fails when the pseudo-invariant features (Rosenfeld, Ravikumar, and Risteski 2020) or geometric skews exist (Nagarajan, Andreassen, and Neyshabur 2021). Under such circumstances, the classifier will utilize both the causal and spurious features, leading to a violation of invariant causal prediction. To address the first issue, we propose an information-theoretical formulation of invariant causal prediction and adopt a variational approximation to ease the optimization procedure. To tackle the second issue, we emphasize that the use of pseudo-invariant features or geometric skews will inevitably increase the mutual information between the inputs and the representations. Thus, to mitigate the impact of pseudo-invariant features and geometric skews, we propose to constrain this mutual information, which naturally leads to a

formulation of information bottleneck. Our empirical results show that the proposed approach can effectively improve the accuracy when the pseudo-invariant features and geometric skews exist.

Contributions: We propose a novel information-theoretic formulation for domain generalization, termed as invariant information bottleneck (IIB). IIB aims at minimizing invariant risks while at the same time mitigating the impact of pseudo-invariant features and geometric skews. Specifically, our contributions can be summarized as follows:

(1) We propose a novel formulation for invariant causal prediction via mutual information. We further adopt variational approximation to develop tractable loss functions for nonlinear classifiers.

(2) To mitigate the impact of pseudo-invariant features and geometric skews, inspired by the information bottleneck principle, we propose to constrain the mutual information between the inputs and the representations. The effectiveness is verified by the synthetic experiments of failure modes (Ahuja et al. 2021; Nagarajan, Andreassen, and Neyshabur 2021), where IIB significantly improves the performance of IRM.

(3) Empirically, we analyze IIB’s performance with extensive experiments on both synthetic and large-scale benchmarks. We show that IIB is able to eliminate the spurious information better than other existing DG methods, and achieves consistent improvements on 7 datasets by 0.7% on DomainBed (Gulrajani and Lopez-Paz 2020).

Related Work

Domain Generalization

Existing methods of DG can be divided into three categories: (1) **Data Manipulation:** Machine learning models typically rely on diverse training data to enhance the generalization ability. Data manipulation/augmentation methods (Nazari and Kovashka 2020; Riemer et al. 2019) aim to increase the diversity of existing training data with operations including flipping, rotation, etc. Domain randomization (Borrego et al. 2018; Yue et al. 2019; Zakharov, Kehl, and Ilıc 2019) provides more complex operations for image data, such as altering the location/texture of objects, replicating and re-sizing objects. In addition, there are some methods (Riemer et al. 2019; Qiao, Zhao, and Peng 2020; Liu et al. 2018; Truong et al. 2019; Zhao et al. 2019b) that exploits generated data samples to enhance the model generalization ability. (2) **Ensemble Learning** methods (Mancini et al. 2018; Segù, Tonioni, and Tombari 2020) assume that any sample in the test domain can be regarded as an integrated sample of the multiple-source domains, so the overall prediction should be inferred by a combination of the models trained on different domains. (3) **Meta-Learning** aims at learning a general model from multiple domains. In terms of domain generalization, MLDG (Li et al. 2018a) divides data from the multiple domains into meta-train and meta-test to simulate the domain shift situation to learn the general representations. In particular, Meta-Reg (Balaji, Sankaranarayanan, and Chellappa 2018) learns a meta-regularizer for the classifier, and Meta-VIB (Du et al. 2020) learns to generate the

weights in the meta-learning paradigm by regularizing the KL divergence between marginal distributions of representations of the same category but from different domains.

Mutual Information-based Domain Adaptation

Domain Adaptation is an important topic in the direction of transfer learning (Long et al. 2015; Ganin et al. 2016; Tzeng et al. 2017; Long et al. 2018; Zhao et al. 2021, 2020c,b; Li et al. 2020a). The mutual information-based approaches have been widely applied in this area. The key idea is to learn a domain-invariant representation that are informative to the label, which can be formulated as (Zhao et al. 2020a; Li et al. 2020b)

$$\max_Z I(Z, Y) - \lambda I(Z, A) \quad (1)$$

where A is the identity of domains, Z denotes the representation, and Y denotes the labels. Commonly adopted implementations of (1) are DANN (Ganin et al. 2017) and CDANN (Long et al. 2018). These implementations are also often adopted in domain generalization as baselines (Gulrajani and Lopez-Paz 2020).

Invariant Risk Minimization

The above approaches enforces the invariance of the learned representations. On the other hand, Invariant Risk Minimization (IRM) suggests the invariance of feature-conditioned label distribution. Specifically, IRM seeks for an invariant causal prediction such that $\mathbb{E}[Y^e | \Phi(X^e)] = \mathbb{E}[Y^{e'} | \Phi(X^{e'})]$, for all $e, e' \in \mathcal{E}$. The objective of IRM is given by

$$\begin{aligned} \min_{\mathbf{w}, \Phi} \sum_{e \in \mathcal{E}_{\text{train}}} R^e(\mathbf{w} \circ \Phi), \\ \text{s.t. } \mathbf{w} \in \underset{\hat{\mathbf{w}}}{\text{argmin}} R^e(\hat{\mathbf{w}} \circ \Phi), \end{aligned}$$

where R^e is the cross-entropy loss for environment e , Φ is the feature extractor and \mathbf{w} is a linear classifier. Note that the above objective is a bilevel optimization and difficult to optimize. Thus, in (Arjovsky et al. 2019), first-order approximation is adopted and the loss function is given by

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{\text{train}}} R^e(\Phi) + \lambda \cdot \|\nabla_{\mathbf{w}|\mathbf{w}=1.0} R^e(\mathbf{w} \circ \Phi)\|, \quad (2)$$

where $\mathbf{w} \in \mathbb{R}$ is a dummy classifier.

Preliminaries

Failure modes of learning invariant representations are well-known in the literature (Zhao et al. 2019a, 2020a). Recently, some works have focused on characterizing the failure modes of IRM as well (Rosenfeld, Ravikumar, and Risteski 2020; Nagarajan, Andreassen, and Neyshabur 2021). As a motivation, we first briefly summarize these negative findings about IRM below.

Pseudo-invariant Features Even in the linear setting, it has been shown that the original IRM formulation (2) cannot truly recover the features that induce invariant causal predictions (Rosenfeld, Ravikumar, and Risteski 2020). Roughly

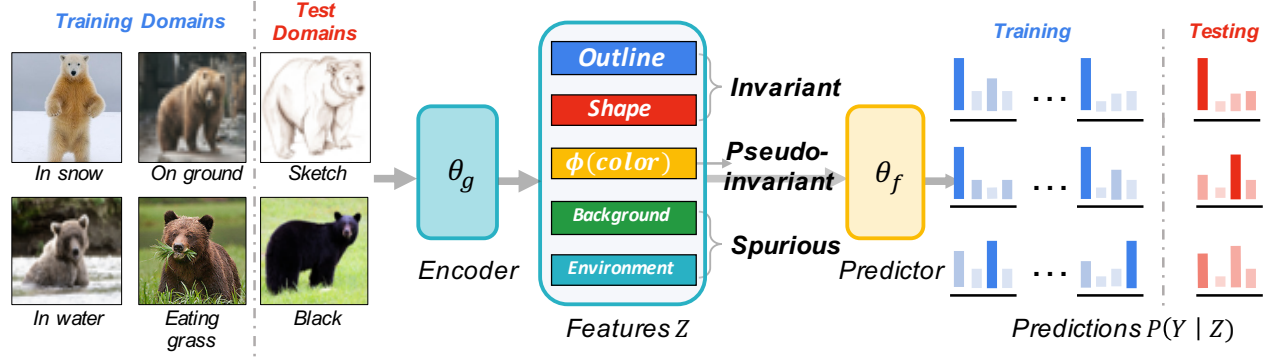


Figure 1: Illustrations of features in OOD generalization. For all the bears in training domains, the predictions $P(Y | Z)$ conditioning on the invariant features (e.g. *outline*) will be correct and invariant, while the predictions conditioning on pseudo-invariant features (possibly *fur color* in this example) are misleading and may affect the generalization ability on test domains. Geometric skews (Nagarajan, Andreassen, and Neyshabur 2021) are the spurious features used as a short-cut for max-margin classifiers. In this example, ERM will use all 5 features as they are informative to labels. IRM, with the invariance constraint, will utilize the first 3 features. IIB, by selecting the minimal sufficient features, only includes the **shape** or **outline**.

speaking, in the linear case, one additional environment could be used to identify one spurious feature, and if the number of environments is smaller than the number of spurious features, some spurious features will leak to the algorithm-recovered causal features, which we call the *pseudo-invariant features*. Specifically, we denote the causal features and spurious features as z_c and z_s respectively. According to the analysis in (Rosenfeld, Ravikumar, and Risteski 2020), there exists a transformation Φ such that $[z_c, \Phi z_s]$ are invariant features across the training dataset. Furthermore, the classifier will utilize $[z_c, \Phi z_s]$ instead of z_c to achieve a lower training error. The OOD generalization may fail due to the inclusion of z_s , which can be arbitrary in the test dataset. An illustration of pseudo-invariant features is shown in Fig. 1.

Geometric Skews The OOD generalization can fail even if we assume the invariant features in the training dataset are also invariant in the test dataset due to the *geometric skews* (Nagarajan, Andreassen, and Neyshabur 2021). It is observed that as the number of training points increase, the ℓ_2 -norm of the max-margin classifier grows. Specifically, we consider the case where an invariant feature z_{inv} is concatenated with a spurious feature z_{sp} such that $\mathbb{P}[z_{sp} \cdot y > 0] > 0.5$. The dataset consists of a majority group S_{maj} where $z_{sp} \cdot y > 0$ (e.g., cows/camels with green/yellow backgrounds) and a minority group S_{min} where $z_{sp} \cdot y < 0$ (e.g., cows/camels with yellow/green backgrounds). Let w_{all} denote the least-norm classifier using invariant features to classify all samples and w_{min} denote the least-norm classifier using invariant features to classify the samples in S_{min} , and we have $\|w_{min}\| \ll \|w_{all}\|$. Hence, the algorithm can use the spurious feature as a short-cut to classify S_{maj} and S_{min} , and then adopt w_{min} to classify the remaining S_{min} . This classifier using spurious feature will have a smaller norm than the invariant classifier, which leads to the failure of OOD generalization.

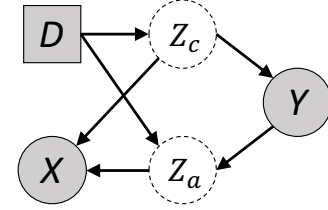


Figure 2: A structural causal model explaining that different parts of an input X have different causal relationships with the model output Y . Observed variables are shaded, while others are with dotted outlines.

Our Proposed Method

In this section, we propose a novel information-theoretic objective of finding invariant causal relationship to overcome the two existing issues in the design of IRM objective.

Invariant Causal Prediction via Mutual Information

Like other casual related works (Chang et al. 2020; Mahajan, Tople, and Sharma 2020), we begin with a structural causal model, shown in Figure 2. For simplicity, we leave out all the unnecessary elements. In general, we can see that an input X can be divided into two variables, the *causal* feature Z_c and *environmental* feature Z_a . In Figure 2, we can readout that both features are correlated with Y , but only Z_c is regarded as a causal feature. Through the concept of d -separation (Pearl 2010), we can readout the conditional independence conditions that all data distributions $\mathcal{P}(D, X, Y)$ should satisfy:

1. $Y \not\perp D$ means the marginal distribution of class label Y can change across domains.
2. $Y \perp\!\!\!\perp D \mid Z_c$ means the class label Y is independent of

domain D conditioned on the causal feature Z_c . The underlying causal mechanism determines that the value of Y comes from its unique causal parent Z_c , which does not change across domains.

3. $Y \not\perp D \mid Z_c, Z_a$ means that the conditional independence will not hold true if conditioned on both the *causal* feature Z_c and the *environmental* features Z_a since Z_a is a collider between D and Y .

The conditional independence tells us that only the real causal relation is stable and remains invariant across domains. In other words, we should eliminate the spurious environmental feature Z_a by seeking the causal feature Z_c that is independent of D from $\Phi(X)$. Particularly, the representation $Z = \Phi(X)$ should have the following two merits: (1) Z does not change among different domains for the same class label Y , hence achieving the conditional invariance of $Y \perp D \mid Z$; (2) Z should be informative of the class label Y (otherwise even a constant $\Phi(\cdot)$ would meet the first goal). The above two conditions coincide with the objective of IRM, and also suggest the following learning objective:

$$\max_{\Phi} I(\Phi(X), Y) - \lambda I(Y, D \mid \Phi(X)), \quad (3)$$

where Φ is the feature extractor.

Proposition 1. Assume $I(Y, D \mid Z) = 0$, then we achieve invariant causal prediction in the sense that $\mathbb{E}[Y \mid \Phi(X) = x, D] = \mathbb{E}[Y \mid \Phi(X) = x]$.

Proof. Note that $I(Y, D \mid Z) = 0$ implies Y and D are independent conditioned on $\Phi(X)$. The conditional independence indicates that $\mathbb{P}(Y \mid \Phi(X) = x, D) = \mathbb{P}(Y \mid \Phi(X) = x)$, thus $\mathbb{E}[Y \mid \Phi(X) = x]$ is fixed and we can achieve invariant causal prediction. ■

On the Failure modes of IRM

In this subsection, we first scrutinize the failure conditions of IRM, i.e., pseudo-invariant features and geometric skews. Based on our analysis, among all the features that satisfy the invariant causal prediction constraint, we propose to use the one with the least capacity, i.e., the one that minimizes $I(X, Z)$. Alternatively, among all the feasible solutions, we are seeking the one that has the largest compression.

With pseudo-invariant features and geometric skews, the failure of existing approaches towards IRM is due to the inclusion of (transformations of) spurious features. We first give an example when the features are one-dimensional and the classifier is linear (Nagarajan, Andreassen, and Neyshabur 2021). Denote the invariant feature, pseudo-invariant feature, feature causing geometric skews, spurious feature as Z_i, Z_p, Z_{sk} , and Z_{sp} . The overall features are $Z = [Z_i, Z_p, Z_{sk}, Z_{sp}]$. In the ERM model, all the features will be adopted and OOD generalization fails. We consider the following optimization problem

$$\begin{aligned} \min_w \sum_{e \in \mathcal{E}_{\text{train}}} R^e(w \cdot Z), \\ \text{s.t. } \|w\|_0 \leq 1, w \in \underset{\hat{w}}{\text{argmin}} R^e(\hat{w} \cdot Z), \end{aligned} \quad (4)$$

where $\|w\|_0 \leq 1$ is the sparsity constraint, and $w \in \underset{\hat{w}}{\text{argmin}} R^e(\hat{w} \cdot Z)$ is the invariant risk constraint of IRM. Due to the sparsity constraint, there are only four choices. Choosing Z_{sp} cannot satisfy the invariant constraint while choosing Z_p or Z_{sk} cannot minimize the empirical risk. Thus, the only optimal solution is $w = [w_1^*, 0, 0, 0]$. Without the sparsity constraint, the optimization problem becomes IRM and Z_i, Z_p, Z_{sk} will be used for classification. Without invariance constraint, Z_{sp} might be chosen as the inclusion of spurious feature can lead to a lower empirical risk.

We then extend this intuition into the loss function design of deep neural networks in the view of mutual information. Suppose Z_1, Z_2 are features extracted from X , we have $I(X, [Z_1, Z_2]) \geq I(X, Z_1)$ as Z_1 is a subset of $[Z_1, Z_2]$. Thus, in order to select the one with the least capacity, we penalize a large $I(X, Z)$ by adding it to the original IRM formulation. To this end, we formulate our objective as

$$\max_{\Phi} I(\Phi(X), Y) - \lambda I(Y, D \mid \Phi(X)) - \beta I(X, \Phi(X)). \quad (5)$$

The term $I(Z, Y) - \beta I(X, Z)$ corresponds to the information bottleneck and $I(Y, D \mid Z)$ implements the IRM principle. As a result, we refer (5) as the *invariant information bottleneck* (IIB) principle.

Loss Function Design

The objective in (5) is still not a tractable loss function as the mutual information of high dimensional vectors is hard to estimate. Similar to VIB (Aleml et al. 2017), we leverage variational approximation to solve this issue. Let $r(z)$ be the approximation to true marginal $p(z)$, and $q(y|z)$ to $p(y|z)$. Meanwhile let $p(z|x)$ be the stochastic encoder. Now the loss function of information bottleneck can be written as

$$\begin{aligned} I(Z, Y) - \beta I(Z, X) \\ \geq \mathbb{E}_{p_{x,y,z}} \left[\log q(y|z) \right] - \beta \mathbb{E}_{p_{x,z}} \left[\log \frac{p(z|x)}{r(z)} \right]. \end{aligned} \quad (6)$$

Optimizing (6) is still a difficult task. Then we transform it with reparametrization operation: We use an encoder of the form $p(z|x; g) = \mathcal{N}(z|g^\mu(x), g^\Sigma(x))$, where g outputs a K -dimensional mean μ of z and a $K \times K$ covariance matrix Σ . Then by the change of variable formula we have $q(z|x)dz = q(\varepsilon)d\varepsilon$, where $z = g(x, \varepsilon)$, $\varepsilon \sim \mathcal{N}(0, 1)$, so we can optimize (6) by optimizing

$$\mathcal{L}_i(g, f_i) + \beta \mathcal{L}_z(g), \quad (7)$$

where $\mathcal{L}_i = \min_{g, f_i} \mathbb{E}_{x,y} [L(y, f_i(g(x)))]$ and $\mathcal{L}_z = \min_g \mathbb{E}_x [KL[q(z|x; g) \| r(z)]]$, where $g(x)$ is the feature extractor, f_i is the classifier, and L is the cross-entropy loss.

We next proceed to deal with $I(Y, D \mid Z)$. Following the rules of variational approximation (Farnia and Tse 2016), we have

$$I(Y, D \mid Z) = H(Y \mid Z) - H(Y \mid D, Z), \quad (8)$$

where $H(Y \mid Z) = -\sup_q \mathbb{E}_{p_{y,z}} [\log q(y|z)]$ and $H(Y \mid D, Z) = -\sup_h \mathbb{E}_{p_{y,z,d}} [\log h(y \mid z, d)]$. Thanks to the universal approximation ability of neural networks, (8) can be

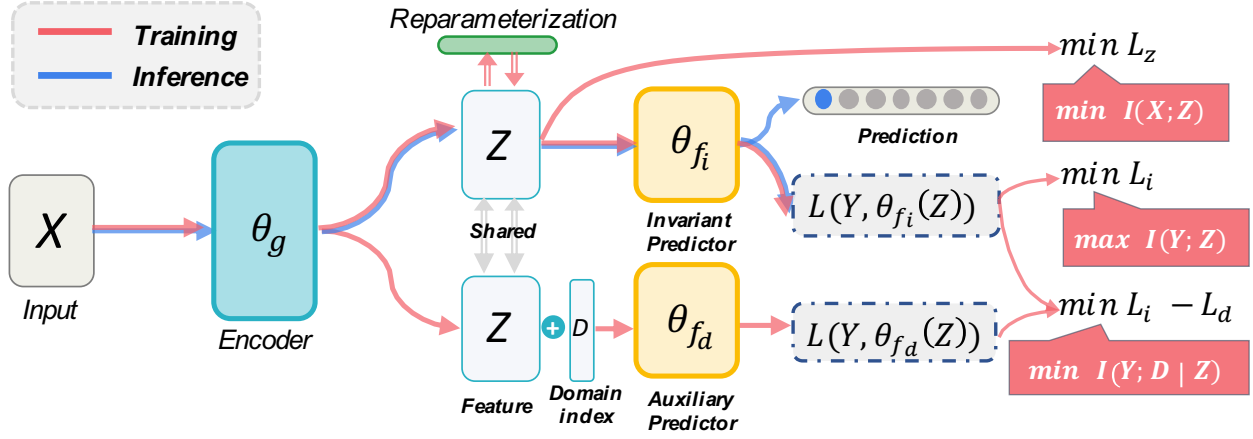


Figure 3: IIB optimizes a model consisting of three parts: (1) an invariant predictor $f_i(Z)$; (2) a domain-dependent predictor $f_d(Z, D)$; (3) an encoder $g(X)$. The three loss terms on the right hand side respectively correspond to the optimization of the three mutual information terms.

Table 1: Accuracy on CS-CMNIST experiment. We split 20% from train set as validation set.

Methods	Validation Acc. (%) \uparrow	Test Acc. (%) \uparrow
ERM (Vapnik 1999)	95.38 \pm 0.03	11.16 \pm 0.31
IRM (Arjovsky et al. 2019)	97.59 \pm 1.39	57.98 \pm 0.86
IB-ERM (Ahuja et al. 2021)	97.64 \pm 0.04	58.47 \pm 0.86
IB-IRM (Ahuja et al. 2021)	97.51 \pm 1.09	71.79 \pm 0.70
IIB ($\lambda = 0$)	92.95 \pm 0.50	69.52 \pm 0.80
IIB ($\beta = 0$)	92.39 \pm 0.50	66.93 \pm 0.33
IIB	98.11 \pm 0.84	74.23 \pm 4.80

written as the subtraction of two classification loss (Farnia and Tse 2016):

$$I(Y, D | Z) = \min_{f_i, g} \underbrace{\mathbb{E}_{x, y} [L(y, f_i(g(x)))]}_{\mathcal{L}_i} - \min_{f_d, g} \underbrace{\mathbb{E}_{x, y, d} [L(y, f_d(g(x), d))]}_{\mathcal{L}_d}, \quad (9)$$

where f_i takes feature z as the input, and $f_d, d = 1, \dots, D$ takes both feature z and domain index d as the input. Overall, we can maximize our IIB objective function by optimizing its tractable lower bound:

$$\min_{g, f_i} \max_{f_d} \mathcal{L}_i(g, f_i) + \beta \mathcal{L}_z(g) + \lambda (\mathcal{L}_i(g, f_i) - \mathcal{L}_d(g, f_d)).$$

Guided by the above objective function, as illustrated in Figure 3, IIB optimizes a model consisting of three parts: (1) an invariant predictor $f_i(Z)$; (2) a domain-dependent predictor $f_d(Z, D)$; (3) an encoder $g(X)$. The code implementation of IIB is released at Github.¹

¹<https://github.com/Luodian/IIB/tree/IIB>

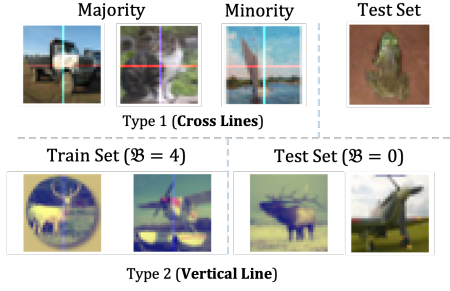
Synthetic Experiments

Experimental Setup

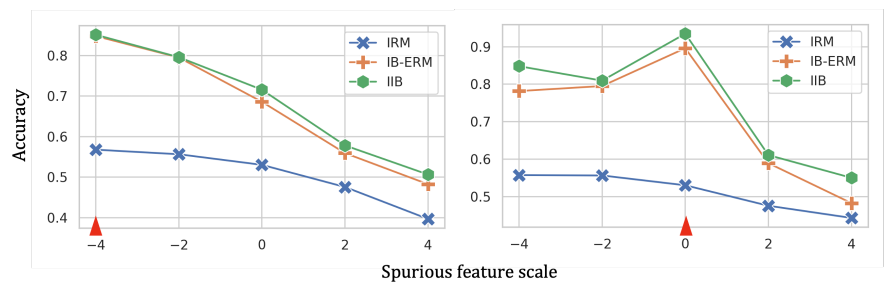
To validate IIB’s efficacy of mitigating the impact of pseudo-invariant features and geometric skews, we adopt two types of synthetic experiments. Both pseudo-invariant features and geometric skews exist in the two experiments.

CS-CMNIST (Ahuja et al. 2021) CS-CMNIST is a ten-way classification task. The images are all drawn from MNIST. There are three environments, two training environments contain each 20,000 images, one test environment contains 20,000 images. There are ten colors associated with ten digit class correspondingly. The probability p_e denotes that the image is colored with associated color. In two training environments, p_e is set to 1 and 0.9, which means the images with certain class are colored with associated color with probability p_e and are colored with random color with probability $1 - p_e$. In test environment, p_e is set to 0, which means all images are colored at random. Overall, the color of images in training domains can be fully predictive to label with spurious features, i.e. using the associated color, but the information disappear at test domain. In CS-CMNIST, if the accuracy drops more at test time, it reflects that relying more on spurious features during training. We will give results of IIB on AC-CMNIST (in DomainBed it’s known as CMNIST) in next section.

Geometric Skew CIFAR10 (Nagarajan, Andreassen, and Neyshabur 2021) There are two types of tasks (as shown in Figure 4 (a)). For the first type, we name it *Cross Lines* experiment, we create ten-valued spurious feature and add a vertical line passing through the middle of each channel, and also a horizontal line through the first channel. For these four lines added, we take the value of $(0.5 \pm 0.5\mathcal{B})$ where $\mathcal{B} \in [-1, 1]$. Four lines, each with 2 choices, then we have a total of $2^4 = 16$ configurations. Among them, we choose the first 10 and denote the 10 configurations to each class in CIFAR10. For i -th configuration, corresponding to i -th class,



(a) Image Examples of two types Geometric Skew CIFAR10



(b) Results of **Vertical Line** experiment.

Figure 4: (a) The above figure represents the image examples in majority/minority group in train set in **Cross Lines** experiment, while colored lines are not included in the test data. The below figure represents the image examples in train/test sets with different spurious feature scales in **Vertical Line** experiment. (b) Accuracy at test domains with different spurious feature scales \mathcal{B} . The upward-pointing red triangle denotes different \mathcal{B} at training domains (we set them to 4 and 0 respectively).

Table 2: Accuracy on Cross-Lines experiment. We split 20% from train set as validation set.

Methods	Validation Acc. (%) \uparrow	Test Acc. (%) \uparrow
ERM (Vapnik 1999)	90.12 \pm 0.12	65.60 \pm 0.27
IRM (Arjovsky et al. 2019)	63.82 \pm 0.25	42.68 \pm 0.32
IB-ERM (Ahuja et al. 2021)	83.93 \pm 0.10	69.70 \pm 0.42
IB-IRM (Ahuja et al. 2021)	81.61 \pm 0.69	65.82 \pm 0.77
IIB ($\lambda = 0$)	79.97 \pm 0.50	69.52 \pm 0.80
IIB ($\beta = 0$)	78.47 \pm 0.50	66.93 \pm 0.33
IIB	92.86 \pm 0.29	71.04 \pm 0.37

we add this line with a probability of $p_{ii} = 0.5$; for other j -th class, we set $p_{ij} = (1 - p_{ii})/10 = 0.05$. Taking the probability means 50% data (the majority group) are correlated with spurious features (the specific colored line corresponding to each class), while other 5% data (the minority group) are correlated with other 9 configurations at random. For the second type, we name it *Vertical Line*, we add a colored line to the last channel of CIFAR10, regardless of the label during training, and vary its brightness during testing. In detail, we add a line with value choose from $\mathcal{B} \in [-4, 4]$. To avoid negative values, all pixels in last channel are added by 4, and then added by \mathcal{B} , and then divided by 9 to make sure the values lie in the range of $[0, 1]$. Such an experiment would artificially create non-orthogonal components, where each data-point is represented on the plane of $(x_{inv}, x_{inv} + x_{env})$, rather than a more easy-to-disentangle representation under (x_{inv}, x_{env}) . As discussed in (Nagarajan, Andreassen, and Neyshabur 2021), the model would be more susceptible to spurious features that may shift during testing.

Observation for results on synthetic experiments

In CS-CMNIST, we compare IIB with several methods, including ERM (Vapnik 1999), IRM, IB-IRM (Ahuja et al. 2021). In particular, IB-IRM (Ahuja et al. 2021) is from a concurrent work, which proposes to combine information bottleneck and IRM to eliminate geometric skews.

Among them (see Table 1), IIB has observable improvements over two synthetic datasets compared with other algorithms. Compare to IB-IRM, which is a direct combination of IB and IRM, our approach took a different approach to optimize the learning objective, which led to further enhancements. In the *Cross Lines* experiment (see Table 2), we train the network on images with colored cross lines (each color corresponds to a specific class in CIFAR10), and test on normal images. From the improvements of IB over IRM, we observe that the information bottleneck structure can help mitigate the failure of IRM in geometric skews. In the *Vertical Line* experiment (see Figure 4 (b)), we train the network on $\mathcal{B} = 4$ or 0, and test on domains with different spurious feature scale \mathcal{B} . The results show that as the offset of spurious feature scale increases, the accuracy of training and testing environments decreases a lot. However, IIB still keeps good results even with large offset, indicating that it’s effectiveness in alleviating the dependence on spurious feature. We have similar observations that information bottleneck (IB) could overcome the geometric skews which fails IRM.

DomainBed Experiments

To empirically corroborate the effectiveness of IIB, we conduct experiments on DomainBed (Gulrajani and Lopez-Paz 2020) with 7 different datasets of different sizes.

Model Selection Strategy We choose two types of model selection strategies out of three in DomainBed. We do not test on the test-domain validation set, since it allows access to test domain while training. During training, the validation set is a subset of training set, we choose the model that performs best on the overall validation set for each domain. This strategy characterizes the in-distribution generalization capability of the model. In leave-one-domain-out cross validation, the training domains are separated from the test domain. This strategy characterizes the out-of-domain distribution generalization capacity of the model. Due to the space limit, we present results on leave-one-domain-out cross validation in Table 3, and put the results on training-domain validation set in supplementary materials.

Table 3: Performance comparison (Acc. %) between the proposed IIB method and the state-of-the-art domain generalization methods with *leave one domain out* model selection strategy. The best accuracy in each dataset is presented in boldface. The average accuracy over all the datasets is also reported.

Methods	Colored-MNIST	Rotated-MNIST	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet	Average
ERM (Vapnik 1999)	36.7 \pm 0.1	97.7 \pm 0.0	77.2 \pm 0.4	83.0 \pm 0.7	65.7 \pm 0.5	41.4 \pm 1.4	40.6 \pm 0.2	63.2
DANN (Ganin et al. 2017)	40.7 \pm 2.3	97.6 \pm 0.2	76.9 \pm 0.4	81.0 \pm 1.1	64.9 \pm 1.2	44.4 \pm 1.1	38.2 \pm 0.2	63.4
CDANN (Li et al. 2018b)	39.1 \pm 4.4	97.5 \pm 0.2	77.5 \pm 0.2	78.8 \pm 2.2	64.3 \pm 1.7	39.9 \pm 3.2	38.0 \pm 0.1	62.2
MLDG (Li et al. 2018a)	36.7 \pm 0.2	97.6 \pm 0.0	77.2 \pm 0.9	82.9 \pm 1.7	66.1 \pm 0.5	46.2 \pm 0.9	41.0 \pm 0.2	64.0
IRM (Arjovsky et al. 2019)	40.3 \pm 4.2	97.0 \pm 0.2	76.3 \pm 0.6	81.5 \pm 0.8	64.3 \pm 1.5	41.2 \pm 3.6	33.5 \pm 3.0	62.0
GroupDRO (Sagawa et al. 2019)	36.8 \pm 0.1	97.6 \pm 0.1	77.9 \pm 0.5	83.5 \pm 0.2	65.2 \pm 0.2	44.9 \pm 1.4	33.0 \pm 0.3	62.7
MMD (Akuzawa, Iwasawa, and Matsuo 2019)	36.8 \pm 0.1	97.8 \pm 0.1	77.3 \pm 0.5	83.2 \pm 0.2	60.2 \pm 5.2	46.5 \pm 1.5	23.4 \pm 9.5	60.7
VREx (Krueger et al. 2020a)	36.9 \pm 0.3	93.6 \pm 3.4	76.7 \pm 1.0	81.3 \pm 0.9	64.9 \pm 1.3	37.3 \pm 3.0	33.4 \pm 3.1	60.6
ARM (Zhang et al. 2020)	36.8 \pm 0.0	98.1 \pm 0.1	76.6 \pm 0.5	81.7 \pm 0.2	64.4 \pm 0.2	42.6 \pm 2.7	35.2 \pm 0.1	62.2
Mixup (Yan et al. 2020)	33.4 \pm 4.7	97.8 \pm 0.0	77.7 \pm 0.6	83.2 \pm 0.4	67.0 \pm 0.2	48.7 \pm 0.4	38.5 \pm 0.3	63.8
RSC (Huang et al. 2020)	36.5 \pm 0.2	97.6 \pm 0.1	77.5 \pm 0.5	82.6 \pm 0.7	65.8 \pm 0.7	40.0 \pm 0.8	38.9 \pm 0.5	62.7
MTL (Blanchard et al. 2021)	35.0 \pm 1.7	97.8 \pm 0.1	76.6 \pm 0.5	83.7 \pm 0.4	65.7 \pm 0.5	44.9 \pm 1.2	40.6 \pm 0.1	63.5
SagNet (Nam et al. 2021)	36.5 \pm 0.1	94.0 \pm 3.0	77.5 \pm 0.3	82.3 \pm 0.1	67.6 \pm 0.3	47.2 \pm 0.9	40.2 \pm 0.2	63.6
IIB(Ours)	39.9 \pm 1.2	97.2 \pm 0.2	77.2 \pm 1.6	83.9 \pm 0.2	68.6 \pm 0.1	45.8 \pm 1.4	41.5 \pm 2.3	64.9

Hyper-parameters and Implementation Details In both selection strategies, for default hyper-parameters (e.g. learning rate, weight decay), we use default settings in DomainBed (e.g. learning rate is set to $1e-3$ for small images and with a selection range of $lr \in [10^{-4.5}, 10^{-2.5}]$). For IIB specific hyper-parameters, we set $\lambda \in [1, 10^2]$, and $\beta \in [10^{-3}, 10^{-4}]$. For backbone feature extractor, in Rotated/Colored-MNIST, we use 4-layers 3x3 ConvNet. For VLCS and PACS, we use ResNet-18 (He et al. 2016). For larger datasets, we opt to ResNet-50. For classifier, we both test linear and non-linear invariant (environment) classifiers. Specifically, in linear classifier, it has only one layer, otherwise it has three MLP layers with two RELU activation layers. For the increased number of parameters in the non-linear classifier, we correspondingly reduce the number of conv-layers in the backbone network to achieve a balance. We test the hyper-paramters and different model implementations on RotatedMNIST, the network is trained for 5000 iterations with batch size set to 128. We report the results in Table 4. We observe that the overall parameters under non-linear classifier setting are not increased too much.

Observation for results on DomainBed

From Table 3, we can see that IIB achieves the best *average* performance on 7 datasets. On the other hand, the results in Table 3 also show that there is no significant advantage of any domain generalization method that can dominate others in small datasets (Colored-MNIST, Rotated-MNIST), which is consistent with the observations in Gulrajani and Lopez-Paz (2020). IIB performs better than others in larger datasets (PACS, Office-Home, DomainNet), hence leading to a better average performance. We opine that the Information Bottleneck is able to better eliminate the noise from the spurious features in large datasets, while when the data set is small, this noise may still be useful as the short-cut in test domain for prediction, thus achieving better results.

Table 4: Different hyper-parameters’ impact to the proposed IIB method on RotatedMNIST with leave-one-domain-out strategy. The results of multiply-add cumulation (MAC) operations and network parameters (Params) are reported.

Classifier Type	MACs	Params	β	λ	Acc. (%) \uparrow
linear	5.83G	370.95K	1e-3	100	61.1
			1e-4	1	94.7
				10	95.3
				100	95.1
non-linear	5.83G	375.33K	1e-3	100	63.2
			1e-4	1	96.8
				10	97.2
				100	97.3

Conclusion

Motivated by the existing limitations of the IRM methods for domain generalization, in this paper we developed a novel information-theoretic approach to overcome these issues. We term our new objective as the invariant information bottleneck (IIB). Our key insight in designing IIB lies in that when the number of training domains is not sufficient to identify all the potential spurious features, we should seek the ones that have the minimum capacity, among all the potential features that satisfy the original IRM objective. To implement IIB, we propose a variational approach to optimize the objective function that goes beyond the previous gradient penalty formulation, which only works for linear classifiers. The superior performance is demonstrated on both synthetic and real datasets through extensive experiments. As a future work, it is interesting to investigate the theoretical foundations of incorporating the information bottleneck principle in nonlinear invariant causal prediction and the effectiveness of IIB on regression tasks.

References

- Ahuja, K.; Caballero, E.; Zhang, D.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. *CoRR*, abs/2106.06607.
- Akuzawa, K.; Iwasawa, Y.; and Matsuo, Y. 2019. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 315–331. Springer.
- Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *CoRR*, abs/1907.02893.
- Balaji, Y.; Sankaranarayanan, S.; and Chellappa, R. 2018. Metareg: Towards domain generalization using meta-regularization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 1006–1016.
- Blanchard, G.; Deshmukh, A. A.; Dogan, Ü.; Lee, G.; and Scott, C. 2021. Domain Generalization by Marginal Transfer Learning. *J. Mach. Learn. Res.*, 22: 2:1–2:55.
- Borrego, J.; Dehban, A.; Figueiredo, R.; Moreno, P.; Bernardino, A.; and Santos-Victor, J. 2018. Applying Domain Randomization to Synthetic Data for Object Category Detection. *CoRR*, abs/1807.09834.
- Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. S. 2020. Invariant Rationalization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, 1448–1458. PMLR.
- Du, Y.; Xu, J.; Xiong, H.; Qiu, Q.; Zhen, X.; Snoek, C. G. M.; and Shao, L. 2020. Learning to Learn with Variational Information Bottleneck for Domain Generalization. In *The 16th European Conference on Computer Vision*, volume 12355 of *Lecture Notes in Computer Science*, 200–216. Springer.
- Farnia, F.; and Tse, D. 2016. A Minimax Approach to Supervised Learning. In *Advances in Neural Information Processing Systems*, 4233–4241.
- Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempit-sky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1): 2096–2030.
- Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempit-sky, V. S. 2017. Domain-Adversarial Training of Neural Networks. In *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, 189–209. Springer.
- Gulrajani, I.; and Lopez-Paz, D. 2020. In Search of Lost Domain Generalization. *CoRR*, abs/2007.01434.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. In *The 14th European Conference on Computer Vision*, volume 9908 of *Lecture Notes in Computer Science*, 630–645. Springer.
- Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging Improves Cross-Domain Generalization. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, 124–140. Springer.
- Krueger, D.; Caballero, E.; Jacobsen, J.; Zhang, A.; Binas, J.; Priol, R. L.; and Courville, A. C. 2020a. Out-of-Distribution Generalization via Risk Extrapolation (REx). *CoRR*, abs/2003.00688.
- Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Priol, R. L.; and Courville, A. 2020b. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*.
- Li, B.; Wang, Y.; Che, T.; Zhang, S.; Zhao, S.; Xu, P.; Zhou, W.; Bengio, Y.; and Keutzer, K. 2020a. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*.
- Li, B.; Wang, Y.; Zhang, S.; Li, D.; Darrell, T.; Keutzer, K.; and Zhao, H. 2020b. Learning Invariant Representations and Risks for Semi-supervised Domain Adaptation. *CoRR*, abs/2010.04647.
- Li, D.; Yang, Y.; Song, Y.; and Hospedales, T. M. 2018a. Learning to Generalize: Meta-Learning for Domain Generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 3490–3497. AAAI Press.
- Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018b. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In *Proceedings of the 15th European Conference on Computer Vision*, 647–663. Springer.
- Liu, A. H.; Liu, Y.; Yeh, Y.; and Wang, Y. F. 2018. A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation. In *Advances in Neural Information Processing Systems*, 2595–2604.
- Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*.
- Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 1640–1650.
- Mahajan, D.; Tople, S.; and Sharma, A. 2020. Domain Generalization using Causal Matching. *CoRR*, abs/2006.07500.
- Mancini, M.; Bulò, S. R.; Caputo, B.; and Ricci, E. 2018. Best Sources Forward: Domain Generalization through Source-Specific Nets. In *2018 IEEE International Conference on Image Processing*, 1353–1357. IEEE.
- Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain Generalization via Invariant Feature Representation. In *Proceedings of the 30th International Conference on Machine Learning*, 10–18. JMLR.org.
- Nagarajan, V.; Andreassen, A.; and Neyshabur, B. 2021. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*.
- Nam, H.; Lee, H.; Park, J.; Yoon, W.; and Yoo, D. 2021. Reducing Domain Gap by Reducing Style Bias. *arXiv:1910.11645*.
- Nazari, N. H.; and Kovashka, A. 2020. Domain Generalization Using Shape Representation. In *Proceedings of European Conference on Computer Vision - ECCV 2020 Workshops*, 666–670. Springer.

- Pearl, J. 2010. Causal Inference. In Guyon, I.; Janzing, D.; and Schölkopf, B., eds., *Causality: Objectives and Assessment (NIPS 2008 Workshop)*, 39–58. JMLR.org.
- Peng, X.; Huang, Z.; Sun, X.; and Saenko, K. 2019. Domain Agnostic Learning with Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, 5102–5112. PMLR.
- Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to Learn Single Domain Generalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 12553–12562. IEEE.
- Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2019. Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference. In *Proceedings of The 7th International Conference on Learning Representations*.
- Rosenfeld, E.; Ravikumar, P.; and Risteski, A. 2020. The Risks of Invariant Risk Minimization. *CoRR*, abs/2010.05761.
- Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *CoRR*, abs/1911.08731.
- Segù, M.; Tonioni, A.; and Tombari, F. 2020. Batch Normalization Embeddings for Deep Domain Generalization. *CoRR*, abs/2011.12672.
- Tachet des Combes, R.; Zhao, H.; Wang, Y.-X.; and Gordon, G. J. 2020. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33.
- Truong, T.; Luu, K.; Duong, C. N.; Le, N.; and Tran, M. 2019. Image Alignment in Unseen Domains via Domain Deep Generalization. *CoRR*, abs/1905.12028.
- Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.
- Vapnik, V. 1999. An overview of statistical learning theory. *IEEE Trans. Neural Networks*, 10(5): 988–999.
- Yan, S.; Song, H.; Li, N.; Zou, L.; and Ren, L. 2020. Improve Unsupervised Domain Adaptation with Mixup Training. *CoRR*, abs/2001.00677.
- Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A. L.; Keutzer, K.; and Gong, B. 2019. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data. In *2019 IEEE/CVF International Conference on Computer Vision*, 2100–2110. IEEE.
- Zakharov, S.; Kehl, W.; and Ilic, S. 2019. Deception-Net: Network-Driven Domain Randomization. In *2019 IEEE/CVF International Conference on Computer Vision*, 532–541. IEEE.
- Zhang, M.; Marklund, H.; Gupta, A.; Levine, S.; and Finn, C. 2020. Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift. *CoRR*, abs/2007.02931.
- Zhao, H.; Dan, C.; Aragam, B.; Jaakkola, T. S.; Gordon, G. J.; and Ravikumar, P. 2020a. Fundamental Limits and Tradeoffs in Invariant Representation Learning. *CoRR*, abs/2012.10713.
- Zhao, H.; Des Combes, R. T.; Zhang, K.; and Gordon, G. 2019a. On Learning Invariant Representations for Domain Adaptation. In *International Conference on Machine Learning*, 7523–7532.
- Zhao, H.; Zhang, S.; Wu, G.; Moura, J. M. F.; Costeira, J. P.; and Gordon, G. J. 2018. Adversarial Multiple Source Domain Adaptation. In *Advances in Neural Information Processing Systems*, 8568–8579.
- Zhao, S.; Li, B.; Xu, P.; and Keutzer, K. 2020b. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*.
- Zhao, S.; Li, B.; Xu, P.; Yue, X.; Ding, G.; and Keutzer, K. 2021. MADAN: multi-source adversarial domain aggregation network for domain adaptation. *International Journal of Computer Vision*, 1–26.
- Zhao, S.; Li, B.; Yue, X.; Gu, Y.; Xu, P.; Hu, R.; Chai, H.; and Keutzer, K. 2019b. Multi-source Domain Adaptation for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, 7285–7298.
- Zhao, S.; Yue, X.; Zhang, S.; Li, B.; Zhao, H.; Wu, B.; Krishna, R.; Gonzalez, J. E.; Sangiovanni-Vincentelli, A. L.; Seshia, S. A.; et al. 2020c. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*.