IMPROVING OOD GENERALIZATION WITH CAUSAL INVARIANT TRANSFORMATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In real-world applications, it is important and desirable to learn a model that performs well on out-of-distribution (OOD) data. Recently, causality has become a powerful tool to tackle the OOD generalization problem, with the core idea resting on the causal mechanism that is invariant across the domains of interest. To leverage the generally unknown causal mechanism, existing works assume the linear form of causal feature or require sufficiently many and diverse training domains, which are usually restrictive in practice. In this work, we obviate these assumptions and tackle the OOD problem without explicitly recovering the causal feature. Our approach is based on transformations that modify the non-causal feature but leave the causal part unchanged, which can be either obtained from prior knowledge or learned from the training data. Under the setting of invariant causal mechanism, we theoretically show that if all such transformations are available, then we can learn a minimax optimal model across the domains using only single domain data. Noticing that knowing a complete set of these causal invariant transformations may be impractical, we further show that it suffices to know only an appropriate subset of these transformations. Based on the theoretical findings, a regularized training procedure is proposed to improve the OOD generalization capability. Extensive experimental results on both synthetic and real datasets verify the effectiveness of the proposed algorithm, even with only a few causal invariant transformations.

1 INTRODUCTION

The success of many machine learning algorithms with empirical risk minimization (ERM) relies on the independent and identically distributed (i.i.d.) hypothesis that training and test data originate from a common distribution. In practice, however, data distributions in different domains or environments are often heterogeneous, due to changing circumstances, selection bias, and time-shifts in the distributions (Meinshausen & Bühlmann, 2015; Rothenhäusler et al., 2021). Accessing data from all the domains of interest, on the other hand, is expensive or even impossible. Consequently, the problem of learning a model that generalizes well on the unseen target distributions is a practically important but also challenging task and has gained much research attention in the past decades (Blanchard et al., 2011; 2021; Fang et al., 2013; Muandet et al., 2013; Volpi et al., 2018).

Since data from some domains are unavailable, assumptions or prior knowledge on the unseen domains are generally required to achieve a guaranteed out-of-distribution (OOD) generalization performance. Recently, causality has become a powerful tool to tackle the OOD problem (Peters et al., 2017; Rojas-Carulla et al., 2018; Arjovsky et al., 2019; Rothenhäusler et al., 2021). This is based on the assumption or observation that the underlying causal mechanism is invariant in general, even though the data distributions may vary with domains. It has been shown that a model would perform well across different domains if such a causal mechanism is indeed captured.

In order to capture the invariant causal mechanism, existing works have assumed a particular form of the causal diagram (Rothenhäusler et al., 2021; Subbaswamy et al., 2019; Mitrovic et al., 2020; Heinze-Deml & Meinshausen, 2021), which is often restrictive in practice and untestable from the observed data. Other works try to recover the so-called "causal feature" from the data to improve the OOD generalization performance (Rojas-Carulla et al., 2018; Chang et al., 2020; Liu et al., 2021; Gimenez & Zou, 2021). These works usually assume a linear form of causal feature (Chang et al.,

2020; Rothenhäusler et al., 2021; Liu et al., 2021; Gimenez & Zou, 2021) or that there are sufficiently many and diverse training domains so that the causal feature could be identified via certain invariant properties (Peters et al., 2016; Rojas-Carulla et al., 2018; Arjovsky et al., 2019). Unfortunately, the linearity assumption is easily violated in real applications like image classification, and it may be expensive or even impossible to ensure that the available domains are indeed sufficient. As such, the identifiability issue of causal feature can hardly be resolved in many real applications.

In this paper, we obviate the aforementioned assumptions and propose a new approach to learn a robust model for OOD generalization under the invariant causal mechanism assumption. We do not try to *explicitly* recover the causal feature; rather, we directly learn a model that takes advantage of the invariant properties. Our approach is based on the observation that though the explicit functional form of the causal feature is generally unknown and maybe also hard to learn, we often have some prior knowledge on the transformations that the causal feature is invariant to, i.e., transformations that modify the input data but do not change their causal features. For example, the shape of the foreground of an image from `MNIST`[1] (LeCun et al., 1998) can be regarded as a causal feature to the task of predicting the digit, while flipping or rotation do not change causal meanings. Henceforth, we refer to these transformations as *causal invariant transformations* (CITs).

Theoretically, we prove that given complete prior knowledge of CITs, it is possible to learn a model with OOD generalization capability using only single domain data. Specifically, we show that if all the CITs are known, then minimizing the loss over all the causally invariant transformed data, which are obtained by applying the CITs to data from the given single domain, would lead to the desired model that achieves minimax optimality across all the domains of interest. However, obtaining all CITs may be impractical. We further show that, for the purpose of OOD generalization, it suffices to know only an appropriate subset of CITs, which is referred to as the causal essential set and is formally defined in Definition 2. Following these theoretical results, we propose to regularize training with the discrepancy between the model outputs of the original data and their transformed versions from the CITs in the causal essential set, to enhance OOD generalization. Empirically, extensive experiments on both synthetic and real-world benchmark datasets, including `PACS` (Li et al., 2017) and `VLCS` (Fang et al., 2013), verify our theoretical findings and demonstrate the effectiveness of the proposed algorithm in terms of OOD performance.

It is worth noting that for the image classification tasks with `PACS` and `VLCS`, we adopt CycleGAN (Zhu et al., 2017) to learn the transformations between pairs of available domains (i.e., styles or backgrounds), which are then used as CITs. The experimental result shows that the proposed regularized training is able to achieve state-of-the-art OOD performance with only a few CITs. We remark that our work is different from the existing domain generation methods that also involve generative methods. For example, Zhou et al. (2020a;b) generate data of inexistent "novel domains", instead of the data from known domains. Also related is Kaushik et al. (2020) that artificially generates the counterfactually-augmented data. Our work is different in that we use CITs to modify the non-causal feature but keep the causal part unchanged. Moreover, in our experiments, these CITs are obtained from prior knowledge or learned from training data, without the need of *human manipulations to each training datum*.

2 Related Work

Since data from some unseen domains are completely unavailable, assumptions or prior knowledge on the data distributions are required to guarantee a good OOD generalization performance. We will briefly review existing domain generalization methods according to these assumptions.

Marginal Transfer Learning A branch of works assume that distributions under different domains are i.i.d. realizations from a superpopulation of distributions and augment the original feature space with the covariate distribution (Blanchard et al., 2011; 2021). This i.i.d. assumption on the data distribution is akin to the random effect model (Laird & Ware, 1982; Bondell et al., 2010) or Bayesian approach (Deely & Lindley, 1981; Ray & van der Vaart, 2020), but may be inappropriate when the difference between domains is irregular, e.g., different styles and backgrounds in the `PACS` and `VLCS` datasets, respectively.

[1] `MNIST` dataset consists of handwritten digits from ten categories.

Robust Optimization Other existing works assume that the considered OOD data are close to the training distribution in terms of a probability distance or divergence, e.g., Wasserstein distance (Sinha et al., 2018; Volpi et al., 2018; Lee & Raginsky, 2018; Yi et al., 2021) and f-divergence (Hu et al., 2018; Gao et al., 2020; Duchi & Namkoong, 2021). They proposed to train the model via distributional robust optimization so that the trained model generalizes well over a set of distributions, the so-called uncertainty set (Ben-Tal et al., 2013; Shapiro, 2017). However, it may be difficult to pick a suitable probability distance and range of uncertainty set in real scenarios (Duchi & Namkoong, 2021). Besides, the distributions in the uncertainty set are actually the distributions of corrupted OOD data (Yi et al., 2021) such as adversarial sample and noisy corrupted data, while the commonly encountered style-transformed OOD data are not included (Hendrycks et al., 2021).

Invariant Feature Another branch of methods make predictions via the features whose (conditional) distributions are invariant across different domains. To this end, they proposed to learn the feature representation by minimizing some loss functions involving domain scatter (Muandet et al., 2013; Ghifary et al., 2016; Li et al., 2018b). Here domain scatter is a quantity characterizing the dissimilarity between (conditional) distributions in different domains, as defined in Ghifary et al. (2016). Li et al. (2018a) and Li et al. (2018c) considered to regularize training to reduce respectively the maximum mean discrepancy of the feature distributions of different domains and the Jensen-Shannon divergence of the feature distributions conditional on the outcome. The rationale behind these methods is to minimize a term that appears in the upper bound of the prediction error in unseen target domains (Ben-David et al., 2007; 2010; Ghifary et al., 2016). Theoretically, the success of these methods hinges on the assumption that other terms in the upper bound are small enough (Ghifary et al., 2016). However, the implication of this assumption is usually unclear and provides little guidance for the practitioner (Chen & Bühlmann, 2020). Although often not stated explicitly, the validity of these methods relies on the *covariate shift* or *label shift* assumption that is implausible if spurious correlations occur under certain domains (Chen & Bühlmann, 2020; Zhou et al., 2021; Kuang et al., 2018; Liu et al., 2021).

Invariant Causal Mechanism As in this paper, many existing works also resort to causality to study the OOD generalization problem (Rojas-Carulla et al., 2018; Arjovsky et al., 2019; Chang et al., 2020; Mitrovic et al., 2020; Ahuja et al., 2020; Liu et al., 2021; Heinze-Deml & Meinshausen, 2021; Gimenez & Zou, 2021). In the last few years, the relationship between causality, prediction, and OOD generalization has gained increasing interest since the seminal work of Peters et al. (2016). The causality-based methods rest on the long-standing assumption that causal mechanism is invariant across different domains (Peters et al., 2017). To utilize the invariant causal mechanism and hence improve OOD generalization, some works impose restrictive assumptions on the causal diagram or structural equations (Subbaswamy et al., 2019; Heinze-Deml & Meinshausen, 2021; Rothenhäusler et al., 2021). Another way is through recovering the causal feature (Rojas-Carulla et al., 2018; Chang et al., 2020; Gimenez & Zou, 2021). For example, Rojas-Carulla et al. (2018) proposed to select causal variables by statistical tests for equality of distributions, and Chang et al. (2020) leveraged some conditional independence relationships induced by the common causal mechanism assumption. It is worth noting that recovering causal feature generally relies on restrictive assumptions, e.g., linear structure model or sufficiently many and diverse training domains (Rojas-Carulla et al., 2018; Chang et al., 2020; Gimenez & Zou, 2021; Peters et al., 2016; Arjovsky et al., 2019; Liu et al., 2021; Krueger et al., 2021). Please see Rosenfeld et al. (2020) for further discussion on these two assumptions. In contrast, our approach relies on a more general causal structural assumption, without the need of restrictive assumptions on the causal diagram, causal features, and training domains.

3 OOD Generalization via Causality

In this section, we consider a more general causal structural model for the OOD generalization problem. We prove that it is feasible to obtain a model with minimax optimality using the causal feature, even if we only have access to the data from a single domain. However, as discussed in the introduction, it may be hard to recover the causal feature exactly. We, therefore, proceed with the aid of CITs and show that a model can still achieve the same guaranteed OOD performance.

3.1 INVARIANT CAUSAL MECHANISM

We begin with a formal definition of the causal structural model used in this paper. In practice, data distributions can vary across domains, but the causal mechanism usually remains unchanged (Peters et al., 2017). We consider the following causal structural model to describe the data generating mechanism:

$$Y = m(g(X), \eta),\ \eta \perp\!\!\!\perp g(X) \text{ and } \eta \sim F, \tag{1}$$

where X, Y are respectively the observed input and the corresponding outcome, $g(X)$ denotes the causal feature, η is some random noise, and $m(\cdot, \cdot)$ represents the unknown structural function. The relationship $\eta \perp\!\!\!\perp g(X)$ means that the noise η is independent of the causal feature $g(X)$, and $\eta \sim F$ indicates that it follows a distribution F that can be unknown.

Remark 1. Notice that the structural model (1) imposes no assumption on the distribution of the input X. Indeed, it describes that in the causal mechanism the outcome Y depends on X *only* through the causal feature $g(X)$. Although the causal feature $g(X)$ is assumed to be independent of noise η, X can correlate with η under a certain domain. There may also exist correlations between causal features and other spurious features, e.g., the correlation between the objective shape and the image background in image classification tasks. Unlike the invariant causal mechanism, these two correlations are supposed to vary across domains and hence are called spurious correlations (Woodward, 2005; Arjovsky et al., 2019). If not treated carefully, the spurious correlations would deteriorate the performance of ERM-based machine learning methods and make the model perform poorly on the target domain (Shen et al., 2020a;b; Liu et al., 2021; Arjovsky et al., 2019). For instance, in an image classification task involving horse and camel, it is very likely that in the training data all the horses are on the grass while the camels are in the desert. The spurious correlation between horse/camel and the background could easily mislead the model to making predictions using the background. Consequently, the trained model would be unreliable on OOD data.

Existing causal learning works consider a similar causal mechanism to (1) but usually impose more structural assumptions, e.g., $g(\cdot)$ is linear and the noise is additive (Peters et al., 2016; Pfister et al., 2019; Rojas-Carulla et al., 2018; Gimenez & Zou, 2021; Liu et al., 2021). Our structural model (1) generalizes existing ones in two ways. First, it allows the form of causal feature $g(\cdot)$ to be nonlinear. Second, the noise η can be non-separable with $g(X)$. Thus, our model constitutes a more flexible construction that is suitable to tasks in which the assumed linear or separable structural models appear implausible, e.g., the image classification task (Arjovsky et al., 2019). Besides, our algorithm proposed in Section 4 does not require *explicitly* learning the causal feature $g(X)$, thus avoid dealing with the identifiability issue of $g(X)$.

3.2 IMPROVING OOD GENERALIZATION VIA CAUSAL FEATURE

Throughout the rest of this paper, we focus on the distributions under structural model (1):

$$\mathcal{P} = \{P_{(X,Y)} \mid (X, Y) \sim P_{(X,Y)} \text{ satisfies structural model (1)}\},$$

with fixed $g(\cdot)$, $m(\cdot, \cdot)$ and F. The goal of this work is to train a model that can generalize well across all the domains following the causal mechanism in structural model (1), i.e., for any distribution $P_{(X,Y)} \in \mathcal{P}$. In particular, we aim to find a model $h^*(\cdot)$ such that

$$h^*(\cdot) \in \mathcal{H}_* := \arg\min_h \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h(X), Y)], \tag{2}$$

where $\mathcal{L}(\cdot, \cdot)$ denotes a loss function, e.g., mean squared error for regression or cross entropy for classification. A similar minimax formulation appears in many existing works for OOD generalization; see, e.g., Rojas-Carulla et al. (2018); Arjovsky et al. (2019); Liu et al. (2021); Bühlmann (2020); Gimenez & Zou (2021), among others.

In contrast with existing methods based on data from sufficiently many domains (Qian et al., 2019; Rojas-Carulla et al., 2018; Sagawa et al., 2020; Liu et al., 2021; Krueger et al., 2021), we next show that if $g(\cdot)$ is known, the goal of learning $h^*(\cdot)$ can be achieved using only single domain data. Let P_{s} be the distribution of the source domain from which the training data are collected. Denote the set of optimal models under P_{s} based on causal feature $g(X)$ by

$$\mathcal{H}_{\mathrm{s}} = \left\{ (\phi \circ g)(\cdot) \;\middle|\; \phi(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w], \quad a.s. \right\}, \tag{3}$$

where $\circ$ is the composition of functions and $a.s.$ stands for "almost surely", i.e., for all w except a set of probability zero. Then we have the following result.

Theorem 1. *If $P_{\mathrm{s}} \in \mathcal{P}$, then $\mathcal{H}_{\mathrm{s}} \subset \mathcal{H}_*$.*

A proof of Theorem 1 can be found in Appendix A.1. Theorem 1 gives a class of models that belong to $\mathcal{H}_*$, the set of solutions to the minimax problem defined in (2). A model in $\mathcal{H}_*$ makes predictions via the causal feature $g(X)$, and can be learned using single domain data if the form of $g(\cdot)$ is known. Theorem 1 generalizes existing results in Rojas-Carulla et al. (2018); Liu et al. (2021), in the sense that it is derived under a more general structural model and also readily includes more loss functions $\mathcal{L}(\cdot, \cdot)$ beyond the mean squared loss.

3.3 OOD Generalization and Causal Invariant Transformation

Theorem 1 shows that it is possible to use only single domain data to learn a class of optimal models $\mathcal{H}_{\mathrm{s}}$ in the minimax sense. However, such a result requires an explicit formulation of causal feature $g(X)$, which is somehow impractical (Arjovsky et al., 2019). On the other hand, learning the causal mechanism from the data may face the issue of identifiability. Thus, in this section, we aim to learn a model of $\mathcal{H}_{\mathrm{s}}$ without the requirement of the explicit form of $g(X)$. The idea of our method is to leverage the transformations that do not change the underlying causal feature.

Specifically, although the explicit form of $g(\cdot)$ is unknown in general, we can have prior knowledge that the causal feature should remain invariant to certain transformations $T(\cdot)$. For example, consider the horse v.s. camel problem in Remark 1. For a given image, the shape of a horse/camel could be the causal feature that determines its category. The *exact* function w.r.t. pixels representing the shape may be hard to obtain. Nevertheless, we do know that the shape does not vary with rotation or flipping. We now formally define these transformations.

Definition 1 (Causal Invariant Transformation (CIT))**.** *A transformation $T(\cdot)$ is called a causal invariant transformation if $(g \circ T)(\cdot) = g(\cdot)$.*

Henceforth, $\mathcal{T}_g = \{T(\cdot) : (g \circ T)(\cdot) = g(\cdot)\}$ denotes the set consisting of all CITs. With $\mathcal{T}_g$ in hand, the following theorem states that $\mathcal{H}_{\mathrm{s}}$ can be obtained by solving a minimax problem constructed from single domain data, even if $g(\cdot)$ is unknown.

Theorem 2. *If $P_{\mathrm{s}} \in \mathcal{P}$, then*

$$\mathcal{H}_{\mathrm{s}} \subset \arg\min_{h} \sup_{T \in \mathcal{T}_g} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(T(X)), Y)], \tag{4}$$

where $\mathcal{H}_{\mathrm{s}}$ is defined in Eq. (3).

A proof of Theorem 2 is given in Appendix A.2. If the minimax optimization problem in (4) has a unique minimum (when, e.g., some convexity conditions on the loss function $\mathcal{L}(\cdot, \cdot)$ hold), Theorem 2 implies that the model performs uniformly well over the transformed data from the transformations in $\mathcal{T}_g$ and further over the distributions in $\mathcal{P}$.

We can also rewrite the minimax problem in (4) as

$$\min_{h} \sup_{P \in \mathcal{P}_{\mathrm{aug}}} \mathbb{E}_{P}[\mathcal{L}(h(X), Y)], \tag{5}$$

where $\mathcal{P}_{\mathrm{aug}} = \{P_{(X',Y)} \mid (X, Y) \sim P_{\mathrm{s}},\ X' = T(X) \text{ with } T \in \mathcal{T}_g\}$. Problem (5) has a similar form to problem (2). Recalling the structural model (1), it can be verified that $\mathcal{P}_{\mathrm{aug}}$ is a subset of $\mathcal{P}$. We then have the following two remarks:

1. $\mathcal{P}_{\mathrm{aug}}$ can be a proper subset of $\mathcal{P}$. Thus, the supremum taken in (5) become more tractable compared with that in (2), as we would require less information of $\mathcal{P}$. To see this, suppose that $(X, \eta) \sim P_{(X,\eta)} = P_X \times F$ for a distribution P_X and that $P_{\mathrm{s}} = P_{(X, m(g(X),\eta))}$. Then for any $P \in \mathcal{P}_{\mathrm{aug}}$, we have $Y \perp\!\!\!\perp X \mid g(X)$ if $(X, Y) \sim P$. However, there can exist $P' \in \mathcal{P}$ so that X is correlated with η and hence the conditional independence no longer holds. Therefore, $P' \in \mathcal{P}$ does not lie in $\mathcal{P}_{\mathrm{aug}}$.
2. In the horse v.s. camel example described in Remark 1, the spurious correlations can result in misleading supervision. The set $\mathcal{P}_{\mathrm{aug}}$, on the other hand, is likely to contain distributions

that do not have these spurious correlations or even entail opposite correlations. Thus, the model that overfits spurious correlations can not generalize well on these distributions. In this case, the model misled by spurious correlations can not be a solution to problem (5).

Although Theorem 2 provides a way to learn a model with guaranteed OOD generalization, it may be computationally hard to calculate the supremum over $\mathcal{T}_g$ when it contains plenty of or possibly infinite transformations. Take image classification tasks for example. Suppose that $\mathcal{T}_g$ contains rotations of θ degree, with $\theta = 1, \ldots, 360$. Computing the loss over a total of 360 transformations is computationally expensive. Thus, it is natural to ask a question: can we substitute $\mathcal{T}_g$ in (4) with a proper subset?

3.4 Learning OOD Generalized Model via Causal Essential Set

In this subsection, we positively answer the question at the end of Section 3.3. We show that it is sufficient to use only a subset of $\mathcal{T}_g$, referred to as *causal essential set*. In the following, we first give a formal definition of causal essential set and then prove that it is indeed the desired subset.

Definition 2 (Causal Essential Set)**.** *For $\mathcal{I}_g \subset \mathcal{T}_g$, $\mathcal{I}_g$ is a causal essential set if for all x_1, x_2 satisfying $g(x_1) = g(x_2)$, there are finite transformations $T_1(\cdot), \cdots, T_K(\cdot) \in \mathcal{I}_g$ such that $(T_1 \circ \cdots \circ T_K)(x_1) = x_2$.*

Clearly, there may be multiple causal essential sets, e.g., $\mathcal{T}_g$ itself is a causal essential set. In most cases, we believe that there exists $\mathcal{I}_g$ that is a proper subset of $\mathcal{T}_g$. For example, rotation with one degree itself forms a causal essential set if $\mathcal{T}_g$ is the set of rotations with $\theta, \theta = 1, 2, \ldots, 360$, degrees.

The next theorem indicates that the prior knowledge on any such causal essential set is sufficient to achieve a guaranteed OOD generalization, using only single domain data. A proof is provided in Appendix B.

Theorem 3. *If $P_{\mathrm{s}} \in \mathcal{P}$, then*

$$\mathcal{H}_{\mathrm{s}} = \arg\min_h \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] \quad \text{subject to } \; h(\cdot) = (h \circ T)(\cdot), \; \forall\, T(\cdot) \in \mathcal{I}_g. \tag{6}$$

where $\mathcal{I}_g$ is any causal essential set of $g(\cdot)$ and $\mathcal{H}_{\mathrm{s}}$ is defined in (3).

Compared with Theorem 2, "$\subset$" related to $\mathcal{H}_{\mathrm{s}}$ in (4) is replaced by "$=$" in this theorem, which provides a stronger theoretical guarantee. Thus, one can also readily obtain the model that generalizes well on OOD data by minimizing the loss w.r.t. any data distribution induced by structural model (1), but require less prior knowledge on CITs. In certain cases, the structure of a causal essential set is simple and it is possible to find this subset of CITs based on prior knowledge. Due to space limit, this is illustrated by an example in Appendix C.1.

4 Algorithm

We now propose an algorithm based on the previous analysis w.r.t. CITs. Let $D(\cdot, \cdot)$ denote some measure of discrepancy satisfying $D(v_1, v_2) = 0$ if $v_1 = v_2$ and $D(v_1, v_2) > 0$ otherwise. Then for any model $h(\cdot)$ and transformation $T(\cdot)$, $\mathbb{E}_{P_{\mathrm{s}}}[D\big(h(X), h(T(X))\big)] = 0$ implies $h(\cdot) = h(T(\cdot))$ almost surely. Together with Theorem 3, we consider the following formulation

$$\min_h \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)], \quad \text{subject to } \; \mathbb{E}_{P_{\mathrm{s}}}\left[\sup_{T \in \mathcal{I}_g} D\big(h(X), h(T(X))\big)\right] = 0,$$

where $\mathcal{I}_g$ is a causal essential set. To obviate the difficulty of solving a constrained optimization problem, we further consider a regularized formulation

$$\min_h \left\{ \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] + \lambda_0 \mathbb{E}_{P_{\mathrm{s}}}\left[\sup_{T \in \mathcal{I}_g} [D(h(X), h(T(X)))]\right] \right\}, \tag{7}$$

with a given regularization constant $\lambda_0 > 0$. Supposing that we have training samples $\{(x_i, y_i)\}_{i=1}^n$, then we propose to minimize the empirical counterpart of (7)

$$\min_h \left\{ \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(T(x_i)), y_i) + \frac{\lambda_0}{n} \sum_{i=1}^n \sup_{T \in \mathcal{I}_g} [D(h(x_i), h(T(x_i)))] \right\}.$$

Algorithm 1 Regularized training with Invariance on Causal Essential set (RICE).

Input: Training set $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, batch size S, learning rate η, training iterations N, model $h_\beta(\cdot)$ with parameter β, initialized parameter β_0, regularization constant λ_0, causal essential set $\mathcal{I}_g$, and discrepancy measure $D(\cdot, \cdot)$.

1: **for** $i = 1, \ldots, n$ **do** ▷ *generate causally invariant transformed samples*
2: Generate transformed samples $\{T(x_i)\}_{T \in \mathcal{I}_g}$.
3: **end for**
4: **for** $t = 0, \cdots, N$ **do** ▷ *minimize the regularized loss*
5: Randomly sample a mini-batch $\mathcal{S} = \{(x_{t_1}, y_{t_1}), \cdots, (x_{t_S}, y_{t_S})\}$ from training set.
6: Fetch the transformed samples $\{T(x_{t_1})\}_{T \in \mathcal{I}_g}, \cdots, \{T(x_{t_S})\}_{T \in \mathcal{I}_g}$.
7: Update model parameters via stochastic gradient descent:
8: $$\beta_{t+1} = \beta_t - \frac{\eta}{S} \sum_{i=1}^{S} \nabla_\beta \mathcal{L}(h_\beta(x_{t_i}), y_{t_i})\Big|_{\beta=\beta_t} + \eta \nabla_\beta \Big\{ \frac{\lambda_0}{S} \sum_{i=1}^{S} \sup_{T \in \mathcal{I}_g} D\big(h_\beta(x_{t_i}), h_\beta(T(x_{t_i}))\big) \Big\}\Big|_{\beta=\beta_t}.$$
9: **end for**

The algorithm of solving (7) via first-order methods is summarized in Algorithm 1, where the update step in line 8 can be readily substituted by other optimization algorithms, e.g., Adam (Kingma & Ba, 2015). We refer to the proposed algorithm as Regularized training with Invariance on Causal Essential set (RICE).

Note that obtaining a complete causal essential set may also be hard in many applications. Nonetheless, we usually have or can learn certain transformations with the desired causal invariance. We will empirically show that the proposed algorithm RICE is able to achieve an improved OOD generalization performance, even with a set of only a few CITs. In this case, we can simply replace $\mathcal{I}_g$ with this set in Algorithm 1.

5 Experiments

In this section, we empirically evaluate the efficacy of the proposed algorithm RICE in real-world datasets. We train the model using data from some of the available domains and evaluate the performance on the data from the rest domains that are not used in training. As suggested in Ye et al. (2021), the OOD data can be classified into two categories, namely, data with correlation shift or with diversity shift. Empirical results show that RICE is able to handle both kinds of OOD data. Due to space limit, part of the empirical results, including a toy experiment of synthetic data mentioned in Section 3.4 and the ablation studies, are provided in Appendix C.1.

5.1 Breaking Spurious Correlation

As we have discussed in Remark 1, the spurious correlation in the data may mislead the model to wrong predictions on OOD data, resulting in correlation shift. In this subsection, we empirically verify that RICE in Algorithm 1 can obviate overfitting such spurious correlations.

Data We use the colored `MNIST` (`C-MNIST`) dataset from Devansh Arpit (2019). As in Devansh Arpit (2019), we vary the colors of both foreground and background of an image.

The original `MNIST` dataset consists of handwritten digits from ten categories, namely, 0 to 9. To construct a training set of `C-MNIST`, we pick two colors for the foreground of the images in a given category, and then randomly replace the foreground color with one of the two colors assigned to the category. The background color of each image is handled similarly. For the test set, we randomly assign colors to the foreground and background of each image from the `MNIST` test set, regardless of its category. Some images from the generated `C-MNIST` dataset are visualized in Figure 4 in Appendix C.4. Construction in this way introduces a spurious correlation between category and color in the training set, but not in the test set. In the following, we will show that the proposed method RICE will not be affected by this spurious correlation.

Setup Our model is a five-layer convolution neural network as also used in Devansh Arpit (2019). For the proposed algorithm RICE, we choose the cross entropy loss for $\mathcal{L}(\cdot, \cdot)$ and the ℓ_2-distance for

Table 1: Accuracy (%) on the C-MNIST test set.

Dataset	ERM	MTL	GroupDRO	DANN	IRM	RICE(OURS)
C-MNIST	13.3	14.7	14.1	28.1	15.8	**96.9**

$D(\cdot, \cdot)$. The model is updated by Adam (Kingma & Ba, 2015), and other hyperparameters are given in Appendix C.2. For a handwritten digit, it is known that the shape of its foreground, rather than the color of either foreground or background, determines its category. Thus, transforming the image background with a color (e.g., black) and its foreground with another color (e.g., white) would be a desired CIT. In our experiment, we simply use the original MNIST images as the transformed data, to show the effectiveness of the proposed regularized training procedure.

As we use the original MNIST dataset in training, the training data can be seen from two domains, i.e., the original MNIST and the C-MNIST datasets. As such, we compare RICE with several widely used domain generalization algorithms using the same training data, including empirical risk minimization (ERM), marginal transfer learning (MTL, Blanchard et al., 2021), group distributionally robust optimization (GroupDRO, Sagawa et al., 2020), domain-adversarial neural networks (DANN, Ganin et al., 2016) and invariant risk minimization (IRM, Arjovsky et al., 2019). See Appendix C.5 for a further introduction of these algorithms. For these baseline algorithms, the hyperparameters are adopted from Gulrajani & Lopez-Paz (2020).

Main Results The empirical results are reported in Table 1. We observe that only the proposed algorithm RICE works well on the OOD data in this experiment. We speculate that this is because, for the baseline algorithms, the misleading supervision signal from the color is memorized by the models, even though the original MNIST images are also included in the training set. However, for RICE, the regularizer penalizes the discrepancy between the model outputs of the colored images and the corresponding MNIST versions, which makes the model insensible to the spurious correlation but more dependent on the invariant causal feature.

5.2 Generalizing on Unseen Domains

In this subsection, we conduct experiments on two benchmark datasets, PACS and VLCS, commonly used in domain generalization. The two datasets correspond to the diversity shift we have mentioned.

Data PACS is an image classification dataset consisting of data from four domains of different styles, i.e., {art, cartoon, photo, sketch}, with seven different categories in each domain. VLCS is a dataset comprised of four photographic domains: {VOC2007, LabelMe, Caltech101, SUN09}, and each domain contains five different categories.

Setup As in Gulrajani & Lopez-Paz (2020), we use the model ResNet50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) as the backbone model, and fine-tune the model with different baseline methods. For RICE, the model is trained by Adam and the used hyperparameters are provided in Appendix C.2. To implement RICE, we need to generate the casually invariant transformed data. In the PACS dataset, each domain represents a style of images, e.g., photo or art. Since varying the style of an image does not change its category, we construct the transformations that modify image styles as CITs. To this end, we use CycleGAN (Zhu et al., 2017) to learn the transformations for each pair of domains in the training set, and then implement RICE using the trained CycleGAN models. In VLCS, the photographic of the image plays a similar role to the style in PACS, and we also apply CycleGAN to learn the transformations. Other generative models may also be used, e.g., StarGAN (Choi et al., 2018), when particularly there are numerous domains.

The procedure of RICE is summarized in Figure 1. We also compare the proposed RICE with other commonly used domain generalization algorithms, as in the previous experiment. Besides, an ablation study is left to Appendix C.3, where we consider only single domain data for training.

Main Results The experimental results on PACS and VLCS are summarized in Tables 2 and 3, respectively. The results of baseline methods are from Gulrajani & Lopez-Paz (2020). It is worth noting that the ERM method here also employs some data augmentation techniques to improve the

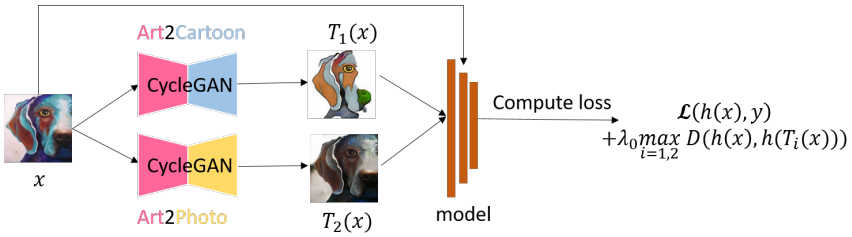


Figure 1: The proposed algorithm RICE on the PACS dataset. The training data are from domains {art, cartoon, photo}, and we would like the model to perform well on the sketch data. This figure describes the training procedure of RICE when an training image is from the art domain.

Table 2: Test accuracy (%) of ResNet50 on the PACS dataset.

Method	P	A	C	S	Avg	Min
ERM	97.2	84.7	80.8	79.3	85.5	79.3
MTL	96.4	87.5	77.1	77.3	84.6	77.1
GroupDRO	96.7	83.5	79.1	78.3	84.4	78.3
DANN	97.3	86.4	77.4	73.5	83.6	73.5
IRM	96.7	84.8	76.4	76.1	83.5	76.1
RICE (OURS)	96.8	87.8	84.3	84.7	**88.4**	**84.3**

Table 3: Test accuracy (%) of ResNet50 on the VLCS dataset.

Method	V	L	C	S	Avg	Min
ERM	74.6	64.3	97.7	73.4	77.5	64.3
MTL	75.3	64.3	97.8	71.5	77.2	64.3
GroupDRO	76.7	63.4	97.3	69.5	76.7	63.4
DANN	77.2	65.1	99.0	73.1	78.6	65.1
IRM	77.3	64.9	98.6	73.4	78.5	64.9
RICE (OURS)	75.1	69.2	98.3	74.6	**79.3**	**69.2**

generalization. The proposed RICE exhibits better OOD generalization capability compared with the baseline methods on the PACS and VLCS datasets, in terms of the average and particularly the worst-case test accuracies. Here we provide an intuitive explanation to the better performance of RICE, using PACS as an example. From Figure 5 in Appendix C.4, we can see that the trained CycleGAN model is likely to introduce spurious correlation with the domains, and that models which capture the spurious correlation would be penalized in RICE. Thus, we believe that RICE can achieve an improved performance because it seeks to make predictions on top of the causal feature (e.g., shape of the object in the images), rather than the spurious feature related to the domains (e.g., style for PACS). Moreover, as also seen from Figure 5, some generated images from CycleGAN are indeed not similar to the original images and are also blurring. However, RICE performs well with these images in the training set, demonstrating the robustness of the proposed training procedure.

6 Concluding Remarks

In this paper, we theoretically show that knowledge of the CITs makes it feasible to learn an OOD generalized model via single domain data. The CITs can be either obtained from prior knowledge or learned from training data, without the need of human manipulations to each training datum. Inspired by our theoretical findings, we propose RICE to achieve an enhanced OOD generalization capability and the effectiveness of RICE is demonstrated empirically over various experiments.

In the experiments, we mainly consider image classification tasks, where the datasets are usually used as benchmarks for domain generalization algorithms. Nevertheless, our theory and the proposed algorithm can apply to other datasets once some CITs are available. For example, in natural language processing (NLP), changing the position of the adverbial or synonym substitution does not change the semantic meaning and hence can be treated as CITs. However, generating such causally invariant sentences via deep learning method is difficult. How to ease the generation and apply the proposed RICE to NLP tasks is left as our future work.

REPRODUCIBILITY STATEMENT

Proofs of the theories in this paper are in Appendix A. Hyperparameters used in experiments in Section 5 are provided in Appendix C.2.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, 2020.

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

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems*, 2011.

Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22:1–55, 2021.

Howard D Bondell, Arun Krishna, and Sujit K Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404 – 426, 2020.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, 2020.

Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. Preprint arXiv:2010.15764, 2020.

Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

J. J. Deely and D. V. Lindley. Bayes empirical bayes. *Journal of the American Statistical Association*, 76(376):833–841, 1981.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Richard Socher Devansh Arpit, Caiming Xiong. Predicting with high correlation features. Preprint arXiv:1910.00164, 2019.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. Preprint, 2020.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1414–1430, 2016.

Jaime Roquero Gimenez and James Zou. Identifying invariant factors across multiple environments with kl regression. Preprint arXiv:2002.08341, 2021.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE International Conference on Computer Vision*, 2021.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, 2018.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation. In *International Conference on Machine Learning*, 2021.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pp. 963–974, 1982.

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

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, 2018.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, 2017.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018a.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Association for the Advancement of Artificial Intelligence*, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision*, 2018c.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, 2021.

Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 2013.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

Qi Qian, Shenghuo Zhu, Jiasheng Tang, Rong Jin, Baigui Sun, and Hao Li. Robust optimization over multiple domains. In *Association for the Advancement of Artificial Intelligence*, 2019.

Kolyan Ray and Aad van der Vaart. Semiparametric bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2020.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. Stable learning via differentiated variable decorrelation. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020a.

Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kunag. Stable learning via sample reweighting. In *Association for the Advancement of Artificial Intelligence*, 2020b.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, 2018.

James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. Preprint arXiv:2106.03721, 2021.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. Improved ood generalization via adversarial training and pre-training. In *International Conference on Machine Learning*, 2021.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Association for the Advancement of Artificial Intelligence*, 2020a.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, 2020b.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. Preprint arXiv:2103.02503, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.

A PROOFS

A.1 PROOF OF THEOREM 1

Restatement of Theorem 1 *If $P_{\mathrm{s}} \in \mathcal{P}$, then $\mathcal{H}_{\mathrm{s}} \subset \mathcal{H}_*$.*

Proof. It suffices to prove that for any $h_{\mathrm{s}} \in \mathcal{H}_{\mathrm{s}}$, we have

$$h_{\mathrm{s}}(\cdot) \in \arg\min_{h} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h(X), Y)]. \tag{8}$$

To prove (8), we only need to show that for any $h(\cdot)$ and $P \in \mathcal{P}$, there exists $Q \in \mathcal{P}$ such that

$$\mathbb{E}_Q[\mathcal{L}(h(X), Y)] \geq \mathbb{E}_P[h_{\mathrm{s}}(X), Y)], \tag{9}$$

and hence

$$\sup_{Q \in \mathcal{P}} \mathbb{E}_Q[\mathcal{L}(h(X), Y)] \geq \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h_{\mathrm{s}}(X), Y)].$$

Recall that

$$\mathcal{H}_{\mathrm{s}} = \left\{ (\phi \circ g)(\cdot) \;\middle|\; \phi(w) \in \arg\min_{z} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w], \quad a.s. \right\}.$$

Since $h_{\mathrm{s}}(\cdot) \in \mathcal{H}_{\mathrm{s}}$, there is some $\phi_{\mathrm{s}}(\cdot)$ satisfying $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot)$ and $\phi_{\mathrm{s}}(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w]$ for almost every w. Suppose $(X, \eta) \sim P_X \times F$ and $(m(g(X), \eta), X) \sim Q$ where P_X is the marginal distributions of X under P. Then

$$\begin{aligned}
\mathbb{E}_Q[\mathcal{L}(h(X), Y) \mid X = x] &= \int_{\mathcal{U}} \mathcal{L}(h(x), m(g(x), u)) P_\eta(du) \\
&\geq \int_{\mathcal{U}} \mathcal{L}(\phi_{\mathrm{s}}(g(x)), m(g(x), u)) P_\eta(du) \\
&= \mathbb{E}_P[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = g(x)] \quad a.s.,
\end{aligned}$$

where the inequality is from the fact

$$\phi_{\mathrm{s}}(w) \in \arg\min_{z} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w] = \arg\min_{z} \int_{\mathcal{U}} \mathcal{L}(z, m(w, u)) P_\eta(du)$$

for almost every w and $\mathcal{U}$ is the support of noise η. Then equation (9) follows by taking expectation and the law of iterated expectation. □

A.2 PROOF OF THEOREM 2

To begin with, we establish two useful lemmas regarding CITs. The first lemma states that $g(\cdot)$ is determined up to an invertible transformation by the transformation that it is invariant to.

For a given function $h(\cdot)$, let $\mathcal{T}_h = \{T(\cdot) : (h \circ T)(\cdot) = h(\cdot)\}$. Then we have the following lemma.

Lemma 1. *For any $h_1(\cdot)$ and $h_2(\cdot)$, $\mathcal{T}_{h_1} \subset \mathcal{T}_{h_2}$ if and only if there exists a function $v(\cdot)$ such that $h_2(\cdot) = (v \circ h_1)(\cdot)$, and $\mathcal{T}_{h_1} = \mathcal{T}_{h_2}$ if and only if there is an invertible function $v(\cdot)$ such that $h_2(\cdot) = (v \circ h_1)(\cdot)$.*

Proof. We only prove the former statement as the latter can be obtained as a corollary of the former. The "if" direction is obvious.

Here we prove the "only if" direction. Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be the range of $h_1(\cdot)$ and $h_2(\cdot)$, respectively. For any $w_1 \in \mathcal{R}_1$ and $w_2 \in \mathcal{R}_2$, define $\mathcal{D}_{h_1, w_1} = \{x : h_1(x) = w_1\}$ and $\mathcal{D}_{h_2, w_2} = \{x : h_2(x) = w_2\}$. Then $h_2(\cdot) = (v \circ h_1)(\cdot)$ if and only if for any $w_2 \in \mathcal{R}_2$, there is some $w_1 \in \mathcal{R}_1$ such that $\mathcal{D}_{h_1, w_1} \subset \mathcal{D}_{h_2, w_2}$. Thus, the former claim holds if we can show the following: $\mathcal{T}_{h_1} \subset \mathcal{T}_{h_2}$ implies that there is some $w_2 \in \mathcal{R}_2$ such that $\mathcal{D}_{h_1, w_1} \subset \mathcal{D}_{h_2, w_2}$ for any $w_1 \in \mathcal{R}_1$. We will prove this by contraction.

Suppose there exists w_1 such that $\mathcal{D}_{h_1, w_1} \not\subset \mathcal{D}_{h_2, w_2}$ for any $w_2 \in \mathcal{R}_2$. Because $\bigcup_{w_2 \in \mathcal{R}_2} \mathcal{D}_{h_2, w_2}$ constitutes the whole space, there is some w_2 such that $\mathcal{D}_{h_1, w_1} \bigcap \mathcal{D}_{h_2, w_2} \neq \varnothing$ and $\mathcal{D}_{h_1, w_1} \not\subset$

$\mathcal{D}_{h_2,w_2}$. Thus, $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2} \neq \varnothing$. Let $x^\dagger$ denote a point in $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2}$ and let x' a point in $\mathcal{D}_{h_2,w_2} \bigcap \mathcal{D}_{h_1,w_1}$. Define T_* as the transformation such that $T_*(x') = x^\dagger$, $T_*(x^\dagger) = x'$ and $T_*(x) = x$ for $x \neq \{x', x^\dagger\}$. Then it is straightforward to verify that $T_* \in \mathcal{T}_{h_1}$ but $T_* \notin \mathcal{T}_{h_2}$, which is a contradiction. □

Thus $g(\cdot)$ can be characterized by $\mathcal{T}_g$ up to an invertible transformation. Define $\mathcal{C}_g = \{g'(\cdot) : g'(\cdot) = (v \circ g)(\cdot)$ for some invertible transformation $v(\cdot)\}$. For any $g' \in \mathcal{C}_g$, by defining

$$\mathcal{H}'_{\mathrm{s}} = \left\{ (\phi \circ g)(\cdot) \;\middle|\; \phi(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g'(X) = w], \quad a.s. \right\},$$

similar arguments as in the proof of Theorem 1 can show $\mathcal{H}'_{\mathrm{s}} \subset \mathcal{H}_*$. To train a model that generalizes well on all the data distributions following the same causal mechanism, any $g'(\cdot) \in \mathcal{C}_g$ is sufficient. Thus, if $\mathcal{T}_g$ is known, to find a model belongs to $\mathcal{H}_*$, one may firstly find an invariant feature map $g'(\cdot)$ such that $\mathcal{T}_{g'} = \mathcal{T}_g$ and then obtain the model according to Theorem 1. However, finding a $g'(\cdot)$ such that $\mathcal{T}_{g'} = \mathcal{T}_g$ is sometimes still a hard task.

For any function $h(\cdot)$, define $\mathcal{I}_h$ in the same way as $\mathcal{I}_g$ with $g(\cdot)$ replaced by $h(\cdot)$ in the definition. We then have the following lemma.

Lemma 2. *For any $h_1(\cdot)$ and $h_2(\cdot)$, if $\mathcal{I}_{h_1} \subset \mathcal{T}_{h_2}$, then there exists a function $v(\cdot)$ such that $h_2(\cdot) = (v \circ h_1)(\cdot)$.*

Proof. Like in the proof of Lemma (1), it suffices to show that $\mathcal{I}_{h_1} \subset \mathcal{T}_{h_2}$ implies for any $w_1 \in \mathcal{R}_1$, there is some $w_2 \in \mathcal{R}_2$ such that $\mathcal{D}_{h_1,w_1} \subset \mathcal{D}_{h_2,w_2}$. We prove this by contraction.

Suppose there is some w_1 such that $\mathcal{D}_{h_1,w_1} \not\subset \mathcal{D}_{h_2,w_2}$ for any $w_2 \in \mathcal{R}_2$. Because $\bigcup_{w_2 \in \mathcal{R}_2} \mathcal{D}_{h_2,w_2}$ is the whole space, there is some w_2 such that $\mathcal{D}_{h_1,w_1} \bigcap \mathcal{D}_{h_2,w_2} \neq \emptyset$ and $\mathcal{D}_{h_1,w_1} \not\subset \mathcal{D}_{h_2,w_2}$. Thus, $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2} \neq \emptyset$. Let $x^\dagger$ be a point in $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2}$ and let x' be a point in $\mathcal{D}_{h_2,w_2} \bigcap \mathcal{D}_{h_1,w_1}$. According to the definition of essential invariant subset, because $h_1(x_1) = h_2(x_2)$, there are finite transformations $T_1(\cdot), \ldots, T_K(\cdot) \in \mathcal{I}_g$ such that $\bar{T}(x') = x^\dagger$ where $\bar{T}(\cdot) = (T_1 \circ \cdots \circ T_K)(\cdot)$. It can be verified that $\mathcal{T}_{h_2}$ is closed with respect to function composition. Hence, $\bar{T}(\cdot) \in \mathcal{T}_{h_2}$. However, $h_2(\bar{T}(x')) = h_2(x^\dagger) \neq w_2 = h_2(x')$, which is a contradiction. □

Restatement of Theorem 2 *If $P_{\mathrm{s}} \in \mathcal{P}$, then*

$$\mathcal{H}_{\mathrm{s}} \subset \arg\min_h \sup_{T \in \mathcal{T}_g} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(T(X)), Y)],$$

where $\mathcal{H}_{\mathrm{s}}$ is defined in (3).

Proof. It suffices to show that for all $h_{\mathrm{s}}(\cdot) \in \mathcal{H}_{\mathrm{s}}$, we have

$$h_{\mathrm{s}}(\cdot) \in \arg\min_h \sup_{T \in \mathcal{T}_g} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(T(X)), Y)]. \tag{10}$$

Note that $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot)$ for some $\phi_{\mathrm{s}}(\cdot)$ and hence is invariant to any transformation $T(\cdot) \in \mathcal{T}_g$. We then have $\sup_{T \in \mathcal{T}_g} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)] = \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(T(X)), Y)]$. Thus, it suffices to prove that for all $h(\cdot)$, there exists $T(\cdot) \in \mathcal{T}_g$ such that

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(T(X)), Y)] \geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)]. \tag{11}$$

According to axiom of choice, there is a choice function a such that $a(w) \in \mathcal{D}_{g,w}$ for almost every w. Define $\tilde{T}$ to be a transformation such that $\tilde{T}(x) = a(w)$ for $x \in \mathcal{D}_{g,w}$. Then $\tilde{T}(\cdot) \in \mathcal{T}_g$ and we have

$$\begin{aligned} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(\tilde{T}(X)), Y) \mid g(X) = w] &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(a(w)), Y) \mid g(X) = w] \\ &\geq \mathbb{E}_{P_{\mathrm{s}}}[\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\ &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s. \end{aligned} \tag{12}$$

By taking expectation on both sides, we can obtain equation (11). □

B PROOF OF THEOREM 3

Restatement of Theorem 3 *If $P_{\mathrm{s}} \in \mathcal{P}$, then*

$$\mathcal{H}_{\mathrm{s}} = \arg\min_{h} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] \quad \text{subject to } \; h(\cdot) = (h \circ T)(\cdot), \; \forall\, T(\cdot) \in \mathcal{I}_g. \tag{13}$$

where $\mathcal{I}_g$ is any causal essential set of $g(\cdot)$ and $\mathcal{H}_{\mathrm{s}}$ is defined in (3).

Proof. We first show

$$\mathcal{H}_{\mathrm{s}} \subset \arg\min_{h} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] \quad \text{subject to } \; h(\cdot) = (h \circ T)(\cdot), \; \forall\, T(\cdot) \in \mathcal{I}_g.$$

Note that the restriction in (13) is equivalent to $\mathcal{I}_g \subset \mathcal{T}_h$. It suffices to show that

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] \geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)] \tag{14}$$

for any $h(\cdot)$ with $\mathcal{I}_g \subset \mathcal{T}_h$ and for any $h_{\mathrm{s}}(\cdot) \in \mathcal{H}_{\mathrm{s}}$. If $\mathcal{I}_g \subset \mathcal{T}_h$, according to Lemma 2, there exists $v(\cdot)$ such that $h(\cdot) = (v \circ g)(\cdot)$. By the definition of $h_{\mathrm{s}}(\cdot)$, there also exists $\phi_{\mathrm{s}}(\cdot)$ satisfying $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot)$ and $\phi_{\mathrm{s}}(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w]$ for almost every w. Thus, we have

$$\begin{aligned}
\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y) \mid g(X) = w] &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(v(w), Y) \mid g(X) = w] \\
&\geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\
&\geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s.
\end{aligned}$$

Then (14) follows by taking expectation.

Next we show the opposite inclusion to prove (13). Suppose $h_*(\cdot)$ is a solution to the optimization problem in (13). Then according Lemma 2, there is some $v_*(\cdot)$ such that $h_*(\cdot) = (v_* \circ g)(\cdot)$. Let $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot) \in \mathcal{H}_{\mathrm{s}}$. Then

$$\begin{aligned}
\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_*(X), Y) \mid g(X) = w] &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(v_*(w), Y) \mid g(X) = w] \\
&\geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\
&= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s.,
\end{aligned} \tag{15}$$

by definition. Because $h_*(\cdot)$ is a solution to the minimization problem, we have

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_*(X), Y)] = \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)].$$

Combining this with (15), we have

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_*(X), Y) \mid g(X) = w] \leq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s. \tag{16}$$

This implies

$$\begin{aligned}
\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(v_*(w), Y) \mid g(X) = w] &\leq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\
&= \min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w] \quad a.s.
\end{aligned}$$

Thus, we conclude that $v_*(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w]$. □

C MORE EXPERIMENTAL RESULTS

C.1 TOY EXAMPLE AND SIMULATION

In the following toy example, we are able to construct an explicit formulation of the causal essential invariant set.

Example 1. Let X be a non-singular 2×2 matrix and $X^{(j)}$ be the j-th column of X for $j = 1, 2$. Suppose that $g(X)$ is the area of the triangle formed by the two points $X^{(1)}$, $X^{(2)}$ and the

origin. Then it is not hard to show that $\{T_{R,\theta}(\cdot), T_{S,a}(\cdot), T_M(\cdot), T_P(\cdot), T_I(\cdot) \mid \theta \in [0, \pi/4],\ a \in [2/3, 3/2]\}$ is an essential invariant set of $g(\cdot)$, where

$$T_{R,\theta}(X) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} X, \quad T_{S,a}(X) = X \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix},$$
$$T_M(X) = X \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad T_P(X) = X \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$
$$T_I(X) = X \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Here $T_{R,\theta}(\cdot)$ rotates the triangle with θ degree clockwise, and $T_{S,a}(\cdot)$ scales the two edges (one connects $X^{(1)}$ to the origin and the other connects $X^{(2)}$ to the origin) of the triangle with a and a^{-1} times, respectively. $T_M(\cdot)$ mirrors the triangle with respect to the x-axis. $T_P(\cdot)$ transforms the triangle to another triangle with same base and height, and $T_I(\cdot)$ transforms the the triangle to another one that is symmetric with respect to the origin. All these transformations are known to keep the triangle area unchanged based on elementary geometry.

Now we verify the effectiveness of the proposed method in the main body using this example.

Data. We consider the following data generation process:

$$\begin{aligned} &X^{(1)} \sim N(0, I_2),\ X^{(2)} \sim N(0, 2I_2),\ X = (X^{(1)}, X^{(2)}), \\ &\epsilon \sim N(0,1),\ \eta = \frac{a\Phi^{-1}(\pi^{-1}\alpha) + \epsilon}{\sqrt{a^2+1}}, \\ &Y = |\det(X)| + \eta, \end{aligned} \tag{17}$$

where I_2 is the identity matrix of order 2, $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. In this data generation process, $|\det(X)|$ is the area of the triangle formed by $X^{(1)}$, $X^{(2)}$ and the origin, and is the causal feature in this example. Here α is the angle between $(X^{(1)} + X^{(2)})/2$ and x-axis, and is correlated with Y in certain domains, with a a parameter that reflects this correlation. However, this correlation is a spurious correlation that changes across domains, i.e., a is set to be different in different domains. In the training population, we pick $a = -3$. We then generate i.i.d. samples of size $1,000$, denoted by $\{(Y_i, X_i)\}_{i=1}^{1000}$, and train a model $h(X, \beta)$ with parameter β to predict Y based on these generated samples.

Model. For any 2×2 matrix

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix},$$

let

$$\begin{aligned} v(X) = (&1, X_{11}, X_{21}, X_{12}, X_{22}, X_{11}^2, X_{21}^2, X_{12}^2, X_{22}^2, \\ &X_{11}X_{21}, X_{11}X_{12}, X_{11}X_{22}, X_{21}X_{12}, X_{21}X_{22}, X_{12}X_{22})^{\mathrm{T}}. \end{aligned}$$

The model is

$$h_\beta(X) = \mathrm{ReLU}(\beta_{[1]}^{\mathrm{T}} v(X)) + \beta_{[2]}^{\mathrm{T}} v(X),$$

where $\beta = (\beta_{[1]}^{\mathrm{T}}, \beta_{[2]}^{\mathrm{T}})^{\mathrm{T}}$ is the model parameter. We pick this model because we have known that $|\det(X)|$ is a function of $v(X)$, and there is some β^* such that $h_{\beta^*}(X) = |\det(X)|$.

Method. Based on the essential invariant set given in Example 1, we define five invariant transformations

$$T_1(X) = \begin{pmatrix} \cos\frac{\pi}{12} & -\sin\frac{\pi}{12} \\ \sin\frac{\pi}{12} & \cos\frac{\pi}{12} \end{pmatrix} X, \quad T_2(X) = X \begin{pmatrix} 1.1 & 0 \\ 0 & 1.1^{-1} \end{pmatrix},$$
$$T_3(X) = X \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad T_4(X) = X \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$
$$T_5(X) = X \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

For ease of notation, we let $T_0(X) = X$ be the identity transformation. We learn the model parameter by minimizing four different loss functions, namely, the empirical risk

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\beta(X_i))^2,$$

the average risk over different transformations

$$\frac{1}{n}\sum_{k=0}^{5}\sum_{i=1}^{n}(Y_i - h_\beta(T_k(X_i)))^2,$$

the maximal risk over different transformations

$$\max_{k=0,\ldots,5}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\beta(T_k(X_i)))^2\right\},$$

and the RICE loss function

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\beta(X_i))^2 + \lambda \max_{k=0,\ldots,5}\left\{\frac{1}{n}\sum_{i=1}^{n}(h_\beta(X_i) - h_\beta(T_k(X_i))^2\right\},$$

where $n = 1000$. In the implementation of RICE, for the given quantities $l_0, \cdots, l_5$, we replace the maximum $\max_{k=0,\ldots,5}\{l_k\}$ in the above losses with the softmax weighting quantity $\sum_{k=0}^{5}\exp(0.2l_k)l_k / \sum_{k=0}^{5}\exp(0.2l_k)$, for ease of computation.

Results. The resulting model is evaluated on i.i.d. sample generated following the data generation process (17) with different a. The following figure plots the squared prediction error of the four methods on test data with different values of a. Each reported value is the average over 200 simulations.

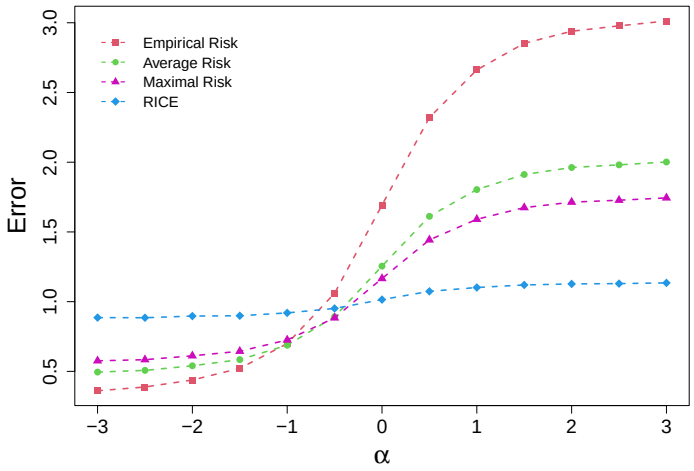


Figure 2: Squared prediction error on test data from distributions with different values of a.

It can be seen that when the test distribution has similar spurious correlations as the training population, minimizing the empirical risk performs the best among the four methods. However, it performs the worst if an opposite spurious correlation appears in the test population. The RICE algorithm has the best worst-case performance, which is consistent with our theoretical analysis. Moreover, the RICE algorithm seems successfully capture the invariant causal mechanism across different environments, as its prediction errors under different test distributions are stable and close to the variance of the intrinsic error η.

Table 4: Hyperparameters of the proposed RICE on C-MNIST, PACS, and VLCS.

Dataset	C-MNIST	PACS	VLCS
Learning Rate	0.1	5e-5	5e-5
Batch Size	128	32	32
Weight Decay	5e-4	0	0
Drop Out	0	0.1	0.1
Epoch	20	20	20
λ_0	0.25	0.5	0.5
β_1	0.9	0.9	0.9
β_2	0.999	0.999	0.999

C.2 HYPERPARAMETERS

We summarize the hyperparameters of the proposed RICE for C-MNIST, PACS, and VLCS datasets in Table 4. The learning rate is decayed by 0.2 at epoch $6, 12$, and 20.

C.3 ABLATION STUDY

In Section 5, for the experiments on PACS and VLCS, we collect training data from several domains for the proposed RICE. However, our theory in Section 3.3 requires only a single domain. Thus, in this subsection, we study the performance of RICE with single domain training data.

Our experiments are conducted on both PACS and VLCS. All the hyperparameters are set to be same with those in Section 5, except the number of training domains—we only use single domain data and hence less training samples for each single experiment. For example, for PACS, if the test domain is sketch, then we run RICE on training data from one of the three other domains (photo, art and cartoon) and report the accuracy on the test domain. To run RICE, the data generated by CycleGAN are used as augmented data and in the regularization term. For a fair comparison, we do not use the CycleGAN that transfer from training domain to test domain and adopt similar experimental settings for ERM.

The results are summarized in Figure 3. We can see that RICE performs much better than the baseline method ERM, which verifies our theoretical conclusions in Theorem 3. Besides, the test accuracy on the target domain can be quite high even when the model is trained using data from a single domain. For example, on VLCS dataset, when test data is from SUN09 domain, the model trained on VOC2007 domain even exhibits a better OOD generalization than the model trained on data from three domains. This implies that, for OOD generalization problem, the number of domains may not be crucial to the performance as long as some representative CITs are available.

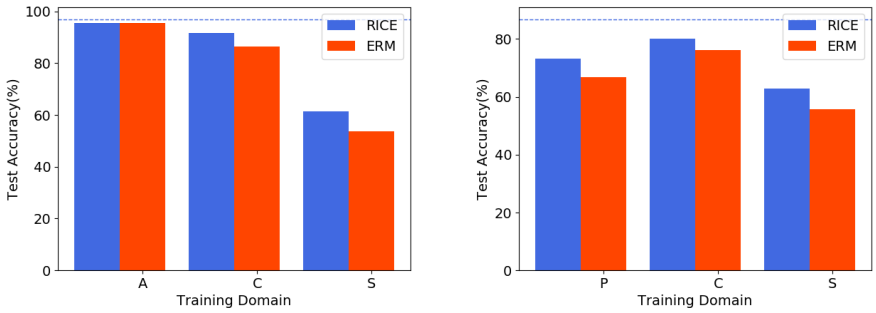
C.4 GENERATED DATA

Our experiments in the main body involve generating causally invariant images. In this subsection, we present visualizations of some generated images for a better understanding of the proposed algorithm.

C-MNIST Figure 4 shows some C-MNIST images. As seen from the training set, there exist spurious correlations between the colors of the foreground or background and the category. However, the correlation disappears in the test set, as the foreground and background colors are randomly assigned.

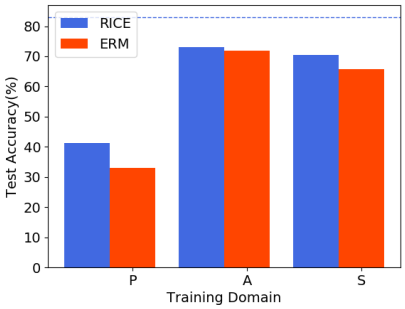
PACS We also present some transformed data from PACS dataset generated by CycleGAN. The CycleGAN is used to simulate CITs as we have clarified the main body of this paper. As the data in PACS come from 7 categories, for each category we pick 4 pictures respectively from domains $\{$photo, art, cartoon, sketch$\}$. The transformed images are shown in Figure 5, where the columns correspond to the styles of $\{$photo, art, cartoon, sketch$\}$, respectively.

Let us look at these generated data over different domains. For the generated images of the photo domain (the first column), the trained CycleGAN tends to alter its color of foreground and add a

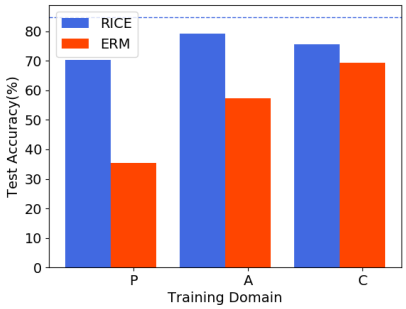


(a) Photo

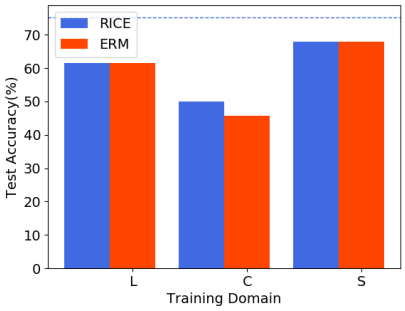
(b) Art



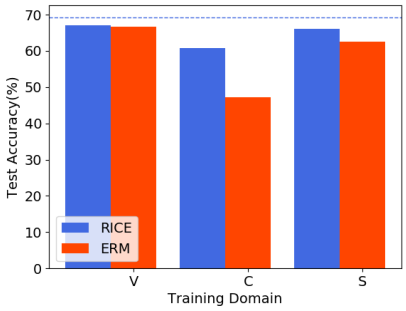
(c) Cartoon



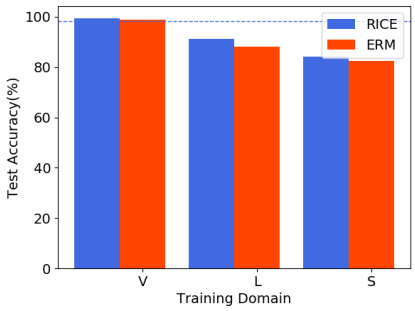
(d) Sketch



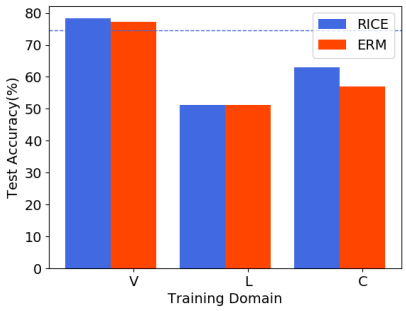
(e) VOC2007



(f) LabelMe



(g) Caltech101



(h) SUN09

Figure 3: Performance of RICE and ERM on the PACS (a-d) and VLCS (e-h) datasets with training data from single domains. Figure title indicates the test domain, and the blue dashed line represents the test accuracy when the training data are from three domains, as reported in Section 5.

(a) Training data in C-MNIST (b) Test data in C-MNIST

Figure 4: Images of the C-MNIST dataset.

background, especially when the original images are from the cartoon and sketch domains. Similar trends exhibit in the generated data of the art domain (the second column). In contrast to the two aforementioned domains, the generated cartoon data in the third column remove the background (if exists) while keep or alter the color of the foreground. The generated sketch data (the fourth column) are more likely to be a grayscale view of the original images. However, for each generated image, the shape of its foreground (i.e., the casual feature to decide the category) does not change when we vary the domains.

The proposed algorithm RICE regularizes the model to encourage the model to be invariant under the CITs, i.e., invariant to the changes of spurious features. This enables the model to be robust to the misleading signal from spurious features and to make predictions via the casual feature. For example, for the dog images in the last row of Figure 5c, which are generated from the images of cartoon style (the third column), the generated dog image of photo style (the first column) has a grass background. However, RICE requires the model to exhibit similar outputs for the two images, hence breaking the spurious correlation between dog and grass.

VLCS Similar to PACS, we present some of the domain transformed data from VLCS dataset generated by CycleGAN. We pick 4 pictures respectively from domains {VOC2007, LabelMe, Caltech101, SUN09} for each of the 5 categories in VLCS. Then we vary the domains of these picked data using the trained CycleGAN models. The transformed data are visualized in Figure 6.

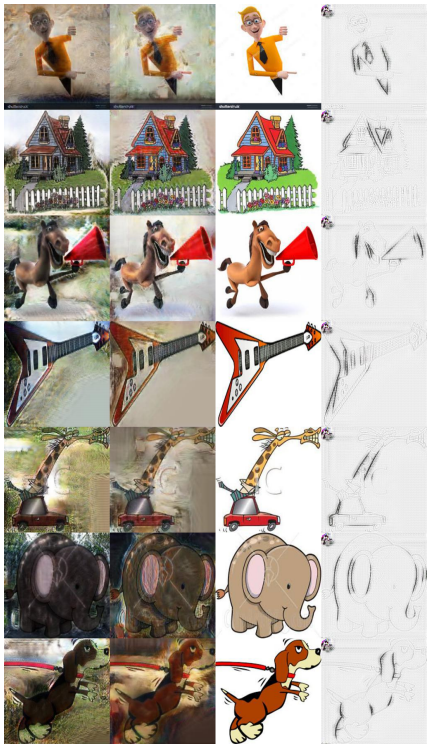
The generated VLCS images exhibit similar behaviors as PACS. Specifically, for a given image from a certain domain, the CycleGAN model tends to deterministically vary the color of the background according to the domains. Thus, the reasoning about the effectiveness of RICE on PACS also applies here.

C.5 Benchmark algorithms

- Empirical Risk minimization (ERM) pools together the data from all the domains and then minimizes the empirical loss to train the model. Notice that here an ImageNET pre-trained model is used.
- Marginal Transfer Learning (MTL, Blanchard et al., 2021) use the mean embedding of the feature distribution in each domain as an input of the classifier.
- Group Distributionally Robust Optimization (GroupDRO, Sagawa et al., 2020) minimizes the largest loss across different domains.
- Domain-Adversarial Neural Networks (DANN, Ganin et al., 2016) use adversarial networks to match the feature distribution in different domains.
- Invariant Risk Minimization (IRM, Arjovsky et al., 2019) learns a feature representation such that the optimal classifiers on top of the representation is the same across the domains.

(a) original domain: photo

(b) original domain: art



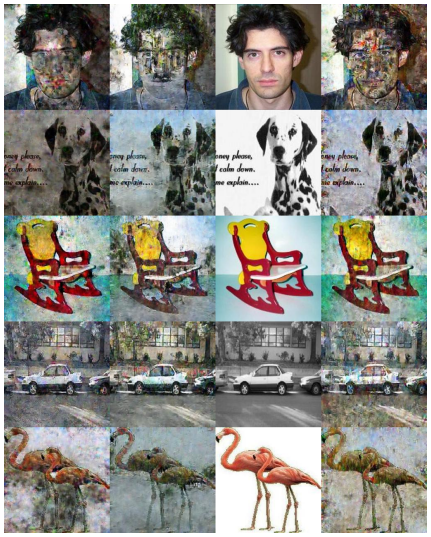
(c) original domain: cartoon

(d) original domain: sketch

Figure 5: Synthetic data of PACS generated by CycleGAN. Columns from left to right correspond to domains of {photo, art, cartoon, sketch}, respectively. Figure title indicates the domain of original data, based on which the data of the rest domains in the figure are generated by CycleGAN.

(a) original domain: VOC2007

(b) original domain: LabelMe



(c) original domain: Caltech101

(d) original domain: SUN09

Figure 6: Synthetic data of `VLCS` generated by CycleGAN. Columns from left to right correspond to domains of {VOC2007, LabelMe, Caltech101, SUN09}, respectively. Figure title indicates the domain of original data, based on which the data of the rest domains in the figure are generated by CycleGAN.