Causally Inspired Regularization Enables Domain General Representations

Olawale Salaudeen
University of Illinois at Urbana-Champaign

Sanmi Koyejo
Stanford University

Abstract

Given a causal graph representing the data-generating process shared across different domains/distributions, enforcing sufficient graph-implied conditional independencies can identify domain-general (non-spurious) feature representations. For the standard input-output predictive setting, we categorize the set of graphs considered in the literature into two distinct groups: (i) those in which the empirical risk minimizer across training domains gives domain-general representations and (ii) those where it does not. For the latter case (ii), we propose a novel framework with regularizations, which we demonstrate are sufficient for identifying domain-general feature representations without a priori knowledge (or proxies) of the spurious features. Empirically, our proposed method is effective for both (semi) synthetic and real-world data, outperforming other state-of-the-art methods in average and worst-domain transfer accuracy.

1 Introduction

A key feature of machine learning is its capacity to generalize across new domains. When these domains present different data distributions, the algorithm must leverage shared structural concepts to achieve out-of-distribution (OOD) or out-of-domain generalization. This capability is vital in numerous important real-world machine learning applications. For example, in safety-critical settings such as autonomous driving, a lack of resilience to unfamiliar distributions could lead to human casualties. Likewise, in the healthcare sector, where ethical considerations are critical, an inability to adjust to shifts in data distribution can result in unfair biases, manifesting as inconsistent performance across different demographic groups.

An influential approach to domain generalization is Invariant Causal Prediction (ICP; [Peters et al., 2016]). ICP posits that although some aspects of data distributions (like spurious or non-causal mechanisms [Pearl, 2010]) may change across domains, certain causal mechanisms remain constant. ICP suggests focusing on these invariant mechanisms for prediction. However, the estimation method for these invariant mechanisms suggested by [Peters et al., 2016] struggles with scalability in high-dimensional feature spaces. To overcome this, Arjovsky et al. [2019] introduced Invariant Risk Minimization (IRM), designed to identify these invariant mechanisms by minimizing an objective. However, requires strong assumptions for identifying the desired domain-general solutions [Ahuja et al., 2021, Rosenfeld et al., 2022]; for instance, observing a number of domains proportional to the spurious features' dimensions is necessary, posing a significant challenge in these high-dimensional settings.

Subsequent variants of IRM have been developed with improved capabilities for identifying domain-general solutions [Ahuja et al., 2020, Krueger et al., 2021, Robey et al., 2021, Wang et al., 2022, Ahuja et al., 2021]. Additionally, regularizers for Distributionally Robust Optimization with subgroup shift have been proposed (GroupDRO) [Sagawa et al., 2019]. However, despite their solid theoretical motivation, empirical evidence suggests that these methods may not consistently deliver domain-general solutions in practice Gulrajani and Lopez-Paz [2020], Kaur et al. [2022], Rosenfeld et al. [2022].

Kaur et al. [2022] demonstrated that regularizing directly for conditional independencies implied by the generative process can give domain-general solutions, including conditional independencies beyond those considered by IRM. However, their experimental approach involves regularization terms that require direct obser-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

vation of spurious features, a condition not always feasible in real-world applications. Our proposed methodology also leverages regularizers inspired by the conditional independencies indicated by causal graphs but, crucially, it does so without necessitating prior knowledge (or proxies) of the spurious features.

1.1 Contributions

In this work,

- we outline sufficient properties to uniquely identify domain-general predictors for a general set of generative processes that include domain-correlated spurious features,
- we propose regularizers to implement these constraints without independent observations of the spurious features, and
- finally, we show that the proposed framework outperforms the state-of-the-art on semi-synthetic and real-world data.

The code for our proposed method is provided at https://github.com/olawalesalaudeen/tcri.

Notation: Capital letters denote random variables, and corresponding lowercase letters denote their value. Unless otherwise stated, we represent latent domain-general features as $Z_{\text{dg}} \in \mathcal{Z}_{\text{dg}} \equiv \mathbb{R}^m$ and spurious latent features as $Z_{\text{spu}} \in \mathcal{Z}_{\text{spu}} \equiv \mathbb{R}^o$. Let $X \in \mathcal{X} \equiv \mathbb{R}^d$ be the observed feature space and the output space of an invertible function $\Gamma : \mathcal{Z}_{\text{dg}} \times \mathcal{Z}_{\text{spu}} \mapsto \mathcal{X}$ and $Y \in \mathcal{Y} \equiv \{0, 1, \ldots, K-1\}$ be the observed label space for a K-class classification task. We then define feature extractors aimed at identifying latent features $\Phi_{\text{dg}} : \mathcal{X} \mapsto \mathbb{R}^m$, $\Phi_{\text{spu}} : \mathcal{X} \mapsto \mathbb{R}^o$ so that $\Phi : \mathcal{X} \mapsto \mathbb{R}^{m+o}$ $\big($that is $\Phi(x) = [\Phi_{\text{dg}}(x); \Phi_{\text{spu}}(x)] \forall x \in \mathcal{X}\big)$. We define e as a discrete random variable denoting domains and $\mathcal{E} = \{P^e(Z_{\text{dg}}, Z_{\text{spu}}, X, Y) : e = 1, 2, \ldots\}$ to be the set of possible domains. $\mathcal{E}_{tr} \subset \mathcal{E}$ is the set of observed domains available during training.

2 Related Work

The source of distribution shift can be isolated to components of the joint distribution. One special case of distribution shift is *covariate shift* [Shimodaira, 2000, Zadrozny, 2004, Huang et al., 2006, Gretton et al., 2009, Sugiyama et al., 2007, Bickel et al., 2009, Chen et al., 2016, Schneider et al., 2020], where only the covariate distribution $P(X)$ changes across domains. Ben-David et al. [2009] give upper-bounds on target error based on the $\mathcal{H}$-divergence between the source and target covariate distributions, which motivates domain alignment methods like the Domain Adversarial Neural Networks [Ganin et al., 2016] and others [Long et al., 2015, Blanchard et al., 2017]. Others have followed up on this work with other notions of covariate distance for domain adaptation, such as mean maximum discrepancy (MMD) [Long et al., 2016], Wasserstein distance [Courty et al., 2017], etc. However, Kpotufe and Martinet [2018] show that these divergence metrics fail to capture many important properties of transferability, such as asymmetry and non-overlapping support. Furthermore, Zhao et al. [2019] shows that even with the alignment of covariates, large distances between label distributions can inhibit transfer; they propose a label conditional importance weighting adjustment to address this limitation. Other works have also proposed conditional covariate alignment [des Combes et al., 2020, Li et al., 2018c,b].

Another form of distribution shift is *label shift*, where only the label distribution changes across domains. Lipton et al. [2018] propose a method to address this scenario. Schrouff et al. [2022] illustrate that many real-world problems exhibit more complex 'compound' shifts than just covariate or label shifts alone.

One can leverage *domain adaptation* to address distribution shifts; however, these methods are contingent on having access to unlabeled or partially labeled samples from the target domain during training. When such samples are available, more sophisticated domain adaptation strategies aim to leverage and adapt spurious feature information to enhance performance Liu et al. [2021], Zhang et al. [2021], Kirichenko et al. [2022]. However, domain generalization, as a problem, does not assume access to such samples [Muandet et al., 2013].

To address the domain generalization problem, Invariant Causal Predictors (ICP) leverage shared causal structure to learn domain-general predictors [Peters et al., 2016]. Previous works, enumerated in the introduction (Section 1), have proposed various algorithms to identify domain-general predictors. Arjovsky et al. [2019]'s proposed invariance risk minimization (IRM) and its variants motivated by domain invariance:

$$\min_{w,\Phi} \frac{1}{|\mathcal{E}_{tr}|} \sum_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi) \text{ s.t. } w \in \operatorname*{argmin}_{\widetilde{w}} R^e(\widetilde{w} \cdot \Phi), \quad \forall e \in \mathcal{E}_{tr},$$

where $R^e(w \circ \Phi) = \mathbb{E}\big[\ell(y, w \cdot \Phi(x))\big]$, with loss function ℓ, feature extractor Φ, and linear predictor w. This objective aims to learn a representation Φ such that predictor w that minimizes empirical risks on average across all domains also minimizes within-domain empirical risk for all domains. However, Rosenfeld et al. [2020], Ahuja et al. [2020] showed that this objective

requires unreasonable constraints on the number of observed domains at train times, e.g., observing distinct domains on the order of the rank of spurious features. Follow-up works have attempted to improve these limitations with stronger constraints on the problem – enumerated in the introduction section.

Our method falls under domain generalization; however, unlike the domain-general solutions previously discussed, our proposed solution leverages different conditions than domain invariance directly, which we show may be more suited to learning domain-general representations.

3 Causality and Domain Generalization

We often represent causal relationships with a causal graph. A *causal graph* is a directed acyclic graph (DAG), $G = (V, E)$, with nodes V representing random variables and directed edges E representing causal relationships, i.e., parents are causes and children are effects. A *structural equation model* (SEM) provides a mathematical representation of the causal relationships in its corresponding DAG. Each variable $Y \in V$ is given by $Y = f_Y(X) + \varepsilon_Y$, where X denotes the parents of Y in G, f_Y is a deterministic function, and ε_Y is an error capturing exogenous influences on Y. The main property we need here is that f_Y is invariant to *interventions* to $V \backslash \{Y\}$ and is consequently invariant to changes in $P(V)$ induced by these interventions. Interventions refer to changes to f_Z, $Z \in V \backslash \{Y\}$.

In this work, we focus on domain-general predictors d_g that are linear functions of features with domain-general mechanisms, denoted as $g_{\text{dg}} \coloneqq w \circ \Phi_{\text{dg}}$, where w is a linear predictor and Φ_{dg} identifies features with domain-general mechanisms. We use domain-general rather than domain-invariant since domain-invariance is strongly tied to the property: $Y \perp\!\!\!\perp e \,|\, Z_{\text{dg}}$ [Arjovsky et al., 2019]. As shown in the subsequent sections, this work leverages other properties of appropriate causal graphs to obtain domain-general features. This distinction is crucial given the challenges associated with learning domain-general features through domain-invariance methods Rosenfeld et al. [2020].

Given the presence of a distribution shift, it's essential to identify some common structure across domains that can be utilized for out-of-distribution (OOD) generalization. For example, Shimodaira [2000] assume $P(Y|X)$ is shared across all domains for the covariate shift problem. In this work, we consider a setting where each domain is composed of observed features and labels, $X \in \mathcal{X}, Y \in \mathcal{Y}$, where X is given by an invertible function Γ of two latent random variables:

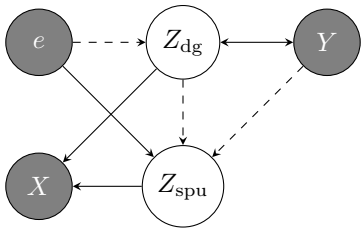


Figure 1: Partial Ancestral Graph representing all non-trivial and valid generative processes (DAGs); dashed edges indicate that an edge may or may not exist.

domain-general $Z_{\text{dg}} \in \mathcal{Z}_{\text{dg}}$ and spurious $Z_{\text{spu}} \in \mathcal{Z}_{\text{spu}}$. By construction, the conditional expectation of the label Y given the domain-general features Z_{dg} is the same across domains, i.e.,

$$\mathbb{E}_{e_i}\left[Y|Z_{\text{dg}} = z_{\text{dg}}\right] = \mathbb{E}_{e_j}\left[Y|Z_{\text{dg}} = z_{\text{dg}}\right] \quad \forall z_{\text{dg}} \in \mathcal{Z}_{\text{dg}}, \forall e_i \neq e_j \in \mathcal{E}. \tag{1}$$

Conversely, this robustness to e does not necessarily extend to spurious features Z_{spu}; in other words, Z_{spu} *may assume values that could lead a predictor relying on it to experience arbitrarily high error rates.* Then, a sound strategy for learning a domain-general predictor – one that is robust to distribution shifts – is to identify the latent domain-general Z_{dg} from the observed features X.

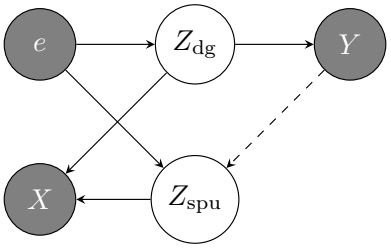
The approach we take to do this is motivated by the *Reichenbach Common Cause Principle*, which claims that if two events are correlated, there is either a causal connection between the correlated events that is responsible for the correlation or there is a third event, a so-called (Reichenbachian) common cause, which brings about the correlation [Hitchcock and Rédei, 2021, Rédei, 2002]. This principle allows us to posit the class of generative processes or causal mechanisms that give rise to the correlated observed features and labels, where the observed features are a function of domain-general and spurious features. We represent these generative processes as causal graphs. Importantly, *the mapping from a node's causal parents to itself is preserved in all distributions generated by the causal graph* (Equation 1), and distributions can vary arbitrarily so long as they preserve the conditional independencies implied by the DAG (Markov Property [Pearl, 2010]).

We now enumerate DAGs that give observe features with spurious correlations with the label.

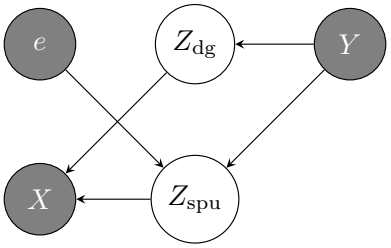
Valid DAGs. We consider generative processes, where both latent features, $Z_{\text{spu}}, Z_{\text{dg}}$, and observed X are correlated with Y, and the observed X is a function of only Z_{dg} and Z_{spu} (Figure 1).

Given this setup, there is an enumerable set of valid

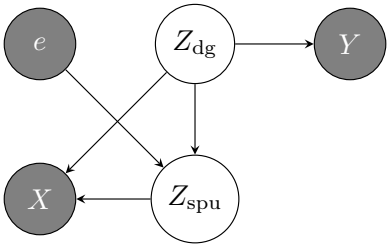
generative processes. Such processes are (i) without cycles, (ii) are feature complete – including edges from Z_{dg} and Z_{spu} to X, i.e., $Z_{\text{dg}} \to X \leftarrow Z_{\text{spu}}$, and (iii) where the observed features mediate domain influence, i.e., there is no direct domain influence on the label $e \not\to Y$. We discuss this enumeration in detail in Appendix B. The result of our analysis is identifying a representative set of DAGs that describe valid generative processes – these DAGs come from orienting the partial ancestral graph (PAG) in Figure 1. We compare the conditional independencies implied by the DAGs defined by Figure 1 as illustrated in Figure 2, resulting in three canonical DAGs in the literature (see Appendix B for further discussion). Other DAGs that induce spurious correlations are outside the scope of this work.



(a) Causal Arjovsky et al. [2019].



(b) Anticausal Rosenfeld et al. [2020].



(c) Fully Informative Causal Ahuja et al. [2021].

Figure 2: **Generative Processes.** Graphical models depicting the structure of possible data-generating processes – shaded nodes indicate observed variables. X represents the observed features, Y represents observed targets, and e represents domain influences (domain indexes in practice). There is an explicit separation of domain-general Z_{dg} and domain-specific Z_{spu} features; they are combined to generate observed X. Dashed edges indicate the possibility of an edge.

Conditional independencies implied by identified DAGs (Figure 2).

Fig. 2a: $\mathbf{Z}_{\text{dg}} \perp\!\!\!\perp \mathbf{Z}_{\text{spu}} \,|\, \{\mathbf{Y}, \mathbf{e}\}$; $Y \perp\!\!\!\perp e \,|\, Z_{\text{dg}}$.

This causal graphical model implies that the mapping from Z_{dg} to its causal child Y is preserved and consequently, Equation 1 holds [Pearl, 2010, Peters et al., 2016]. As an example, consider the task of predicting the spread of a disease. Features may include causes (vaccination rate and public health policies) and effects (public behavior changes such as increased mask-wearing or social distancing).

Fig. 2b: $\mathbf{Z}_{\text{dg}} \perp\!\!\!\perp \mathbf{Z}_{\text{spu}} \,|\, \{\mathbf{Y}, \mathbf{e}\}$; $Z_{\text{dg}} \perp\!\!\!\perp Z_{\text{spu}} \,|\, Y$; $Y \perp\!\!\!\perp e \,|\, Z_{\text{dg}}$, $Z_{\text{dg}} \perp\!\!\!\perp e$.

The causal graphical model does not directly imply that $Z_{\text{dg}} \to Y$ is preserved across domains. However, in this work, it represents the setting where the inverse of the causal direction is preserved (inverse: $Z_{\text{dg}} \to Y$), and thus Equation 1 holds. A context where this setting is relevant is in healthcare where medical conditions (Y) cause symptoms (Z_{dg}), but the prediction task is often predicting conditions from symptoms, and this mapping $Z_{\text{dg}} \to Y$, opposite of the causal direction, is preserved across distributions.

Fig. 2c: $Y \perp\!\!\!\perp e \,|\, Z_{\text{dg}}$; $Z_{\text{dg}} \perp\!\!\!\perp e$.

Similar to Figure 2a, this causal graphical model implies that the mapping from Z_{dg} to its causal child Y is preserved, so Equation 1 holds [Pearl, 2010, Peters et al., 2016]. This setting is especially interesting because it represents a **Fully Informative Invariant Features setting**, that is $Z_{\text{spu}} \perp\!\!\!\perp Y \,|\, Z_{\text{dg}}$ [Ahuja et al., 2021]. As an example of this, we can consider the task of predicting hospital readmission rates. Features may include the severity of illness, which is a direct cause of readmission rates, and also include the length of stay, which is also caused by the severity of illness. However, length of stay may not be a cause of readmission; the correlation between the two would be a result of the confounding effect of a common cause, illness severity.

We call the condition $\mathbf{Y} \perp\!\!\!\perp \mathbf{e} \,|\, \mathbf{Z}_{\text{dg}}$ the *domain invariance property.* This condition is common to all the DAGs in Figure 2. We call the condition $Z_{\text{dg}} \perp\!\!\!\perp Z_{\text{spu}} \,|\, \{Y, e\}$ the *target conditioned representation independence (TCRI) property.* This condition is common to the DAGs in Figure 2a, 2b. In the settings

Table 1: Generative Processes and Sufficient Conditions for Domain-Generality

	Graphs in Figure 2		
	(a)	(b)	(c)
$Z_{\text{dg}} \perp\!\!\!\perp Z_{\text{spu}} \mid \{Y, e\}$	✓	✓	✗
Identifying Z_{dg} is necessary	✓	✓	✗

considered in this work, the TCRI property is equivalently $\mathbf{Z}_{\text{dg}} \perp\!\!\!\perp \mathbf{Z}_{\text{spu}} \mid \mathbf{Y} \forall \mathbf{e} \in \mathcal{E}$ since e will simply index the set of empirical distributions available at training.

Domain generalization with conditional independencies. Kaur et al. [2022] showed that sufficiently regularizing for the correct conditional independencies described by the appropriate DAGs can give domain-general solutions, i.e., identifies Z_{dg}. However, in practice, one does not (partially) observe the latent features independently to regularize directly. Other works have also highlighted the need to consider generative processes when designing robust algorithms to distribute shifts [Veitch et al., 2021, Makar et al., 2022]. However, previous work has largely focused on regularizing for the domain invariance property, ignoring the conditional independence property $Z_{\text{dg}} \perp\!\!\!\perp Z_{\text{spu}} \mid \{Y, e\}$.

Sufficiency of ERM under Fully Informative Invariant Features. Despite the known challenges of learning domain-general features from the domain-invariance properties in practice, this approach persists, likely due to it being the only property shared across all DAGs. We alleviate this constraint by observing that Graph (Fig. 2c) falls under what Ahuja et al. [2021] refer to as the fully informative invariant features settings, meaning that Z_{spu} is redundant, having only information about Y that is already in Z_{dg}. Ahuja et al. [2021] show that the empirical risk minimizer is domain-general for bounded features.

Easy vs. hard DAGs imply the generality of TCRI. Consequently, we categorize the generative processes into *easy* and *hard* cases Table 1: *(i) easy* meaning that minimizing average risk gives domain-general solutions, i.e., ERM is sufficient (Fig. 2c), and *(ii) hard* meaning that one needs to identify Z_{dg} to obtain domain-general solutions (Figs. 2a-2b). We show empirically that regularizing for $Z_{\text{dg}} \perp\!\!\!\perp Z_{\text{spu}} \mid Y \forall e \in \mathcal{E}$ also gives a domain-general solution in the easy case. **The generality of TCRI** follows from its sufficiency for identifying domain-general Z_{dg} in the hard cases while still giving domain-general solutions empirically in the easy case.

4 Proposed Learning Framework

We have now clarified that hard DAGs (i.e., those not solved by ERM) share the TCRI property. The challenge is that Z_{dg} and Z_{spu} are not independently observed; otherwise, one could directly regularize. Existing work such as Kaur et al. [2022] empirically study semi-synthetic datasets where Z_{spu} is (partially) observed and directly learn Z_{dg} by regularizing that $\Phi(X) \perp\!\!\!\perp Z_{\text{spu}} \mid Y, e$ for feature extractor Φ. To our knowledge, we are the first to leverage the TCRI property without requiring observation of Z_{spu}. Next, we set up our approach with some key assumptions. The first is that the observed distributions are Markov to an appropriate DAG.

Assumption 4.1. All distributions, sources and targets, are generated by one of the structural causal models $\mathcal{SCM}$ that follow:

$$\overbrace{\mathcal{SCM}(e)}^{causal} := \begin{cases} Z_{\text{dg}}^{(e)} \sim P_{Z_{\text{dg}}}^{(e)}, \\ Y^{(e)} \leftarrow \langle w_{\text{dg}}^*, Z_{\text{dg}}^{(e)} \rangle + \eta_Y, \\ Z_{\text{spu}}^{(e)} \leftarrow \langle w_{\text{spu}}^*, Y \rangle + \eta_{Z_{\text{spu}}}^{(e)}, \\ X \leftarrow \Gamma(Z_{\text{dg}}, Z_{\text{spu}}), \end{cases} \tag{2}$$

$$\overbrace{\mathcal{SCM}(e)}^{anticausal} := \begin{cases} Y^{(e)} \sim P_Y, \\ Z_{\text{dg}}^{(e)} \leftarrow \langle \widetilde{w}_{\text{dg}}, Y \rangle + \eta_{Z_{\text{dg}}}^{(e)}, \\ Z_{\text{spu}}^{(e)} \leftarrow \langle w_{\text{spu}}^*, Y \rangle + \eta_{Z_{\text{spu}}}^{(e)}, \\ X \leftarrow \Gamma(Z_{\text{dg}}, Z_{\text{spu}}), \end{cases} \tag{3}$$

or

$$\overbrace{\mathcal{SCM}(e)}^{FIIF} := \begin{cases} Z_{\text{dg}}^{(e)} \sim P_{Z_{\text{dg}}}^{(e)}, \\ Y^{(e)} \leftarrow \langle w_{\text{dg}}^*, Z_{\text{dg}}^{(e)} \rangle + \eta_Y, \\ Z_{\text{spu}}^{(e)} \leftarrow \langle w_{\text{spu}}^*, Z_{\text{dg}} \rangle + \eta_{Z_{\text{spu}}}^{(e)}, \\ X \leftarrow \Gamma(Z_{\text{dg}}, Z_{\text{spu}}), \end{cases} \tag{4}$$

where $P_{Z_{\text{dg}}}$ is the causal covariate distribution, w's are linear generative mechanisms, η's are exogenous independent noise variables, and $\Gamma : \mathcal{Z}_{\text{dg}} \times \mathcal{Z}_{\text{spu}} \to \mathcal{X}$ is an invertible function. It follows from having causal mechanisms that we can learn a predictor w_{dg}^* for Z_{dg} that is domain-general (Equation 2-4) – w_{dg}^* inverts the mapping $\widetilde{w}_{\text{dg}}$ in the anticausal case.

These structural causal models (Equation 2-4) correspond to causal graphs Figures 2a-2c, respectively.

Assumption 4.2 (Structural)**.** Causal Graphs and their distributions are Markov and Faithful [Pearl, 2010].

Given Assumption 4.2, we aim to leverage TCRI property ($Z_{\text{dg}} \perp\!\!\!\perp Z_{\text{spu}} \mid Y \forall e \in \mathcal{E}_{tr}$) to learn the latent Z_{dg} without observing Z_{spu} directly. We do this by learning two feature extractors that, together, recover Z_{dg} and Z_{spu} and satisfy TCRI (Figure 3). We formally define these properties as follows.

Definition 4.3 (Total Information Criterion (TIC))**.** $\Phi = \Phi_{\text{dg}} \oplus \Phi_{\text{spu}}$ satisfies TIC with respect to random

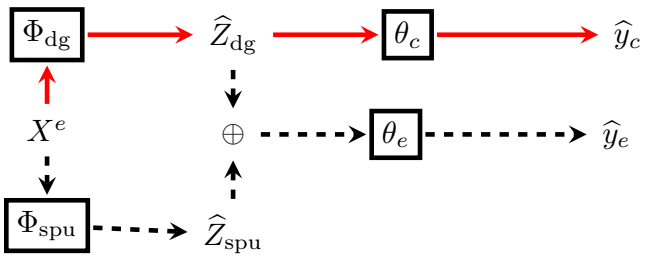


Figure 3: **Modeling approach.** During training, both representations, Φ_{dg}, and Φ_{spu}, generate domain-general and domain-specific predictions, respectively. However, only the domain-invariant representations/predictions are used during testing – indicated by the solid red arrows.

variables X, Y, e if for $\Phi(X^e) = [\Phi_{\text{dg}}(X^e); \Phi_{\text{spu}}(X^e)]$, there exists a linear operator $\mathcal{T}$ s.t., $\mathcal{T}(\Phi(X^e)) = [Z^e_{\text{dg}}; Z^e_{\text{spu}}] \forall e \in \mathcal{E}_{tr}$.

In other words, a feature extractor that satisfies the total information criterion recovers the complete latent feature sets Z_{dg}, Z_{spu}. This allows us to define the proposed implementation of the TCRI property non-trivially – the conditional independence of subsets of the latents may not have the same implications on domain generalization. We note that $X \perp\!\!\!\perp Y | Z_{\text{dg}}, Z_{\text{spu}}$, so X has no information about Y that is not in $Z_{\text{dg}}, Z_{\text{spu}}$.

Definition 4.4 (Target Conditioned Representation Independence)**.** $\Phi = \Phi_{\text{dg}} \oplus \Phi_{\text{spu}}$ satisfies TCRI with respect to random variables X, Y, e if $\Phi_{\text{dg}}(X) \perp\!\!\!\perp \Phi_{\text{spu}}(X) \,|\, Y \forall e \in \mathcal{E}$.

Proposition 4.5. *Assume that $\Phi_{dg}(X)$ and $\Phi_{spu}(X)$ are correlated with Y. Given Assumptions 4.1-4.2 and a representation $\Phi = \Phi_{dg} \oplus \Phi_{spu}$ that satisfies TIC, $\Phi_{dg}(X) = Z_{dg} \iff \Phi$ satisfies TCRI. (see Appendix C for proof).*

Proposition 4.5 shows that TCRI is necessary and sufficient to identify Z_{dg} from a set of training domains. We note that we can verify if $\Phi_{\text{dg}}(X)$ and $\Phi_{\text{spu}}(X)$ are correlated with Y by checking if the learned predictors are equivalent to chance. Next, we describe our proposed algorithm to implement the conditions to learn such a feature map. Figure 3 illustrates the learning framework.

Learning Objective: The first term in our proposed objective is

$$\mathcal{L}_{\Phi_{\text{dg}}} = \mathcal{R}^e(\theta_c \circ \Phi_{\text{dg}}),$$

where $\Phi_{\text{dg}} : \mathcal{X} \mapsto \mathbb{R}^m$ is a feature extractor, $\theta_c : \mathbb{R}^m \mapsto \mathcal{Y}$ is a linear predictor, and $\mathcal{R}^e(\theta_c \circ \Phi_{\text{dg}}) = \mathbb{E}\big[\ell(y, \theta_c \cdot \Phi(x))\big]$ is the empirical risk achieved by the feature extractor and predictor pair on samples from domain e. Φ_{dg} and θ_c are designed to capture the domain-general portion of the framework.

Next, to implement the total information criterion, we use another feature extractor $\Phi_{\text{spu}} : \mathcal{X} \mapsto \mathbb{R}^o$, designed to capture the domain-specific information in X that is not captured by Φ_{dg}. Together, we have $\Phi = \Phi_{\text{dg}} \oplus \Phi_{\text{spu}}$ where Φ has domain-specific predictors $\theta_e : \mathbb{R}^{m+o} \mapsto \mathcal{Y}$ for each training domain, allowing the feature extractor to utilize domain-specific information to learn distinct optimal domain-specific (non-general) predictors:

$$\mathcal{L}_{\Phi} = \mathcal{R}^e\big(\theta_e \circ \Phi\big).$$

$\mathcal{L}_{\Phi}$ aims to ensure that Φ_{dg} and Φ_{spu} capture all of the information about Y in X – total information criterion. Since we do not know o, m, we select them to be the same size on our experiments; o, m could be treated as hyperparameters though we do not treat them as such.

Finally, we implement the TCRI property (Definition 4.4). We denote $\mathcal{L}_{TCRI}$ to be a conditional independence penalty for Φ_{dg} and Φ_{spu}. We utilize the Hilbert Schmidt independence Criterion (HSIC) [Gretton et al., 2007] as $\mathcal{L}_{TCRI}$. However, in principle, any conditional independence penalty can be used in its place. **HSIC:**

$$\begin{aligned}
&\mathcal{L}_{TCRI}(\Phi_{\text{dg}}, \Phi_{\text{spu}}) \\
&= \frac{1}{2} \sum_{k \in \{0,1\}} \widehat{HSIC}\Big(\Phi_{\text{dg}}(X), \Phi_{\text{spu}}(X)\Big)^{y=k} \\
&= \frac{1}{2} \sum_{k \in \{0,1\}} \frac{1}{n_k^2} \text{tr}\Big(\mathbf{K}_{\Phi_{\text{dg}}} \mathbf{H}_{n_k} \mathbf{K}_{\Phi_{\text{spu}}} \mathbf{H}_{n_k}\Big)^{y=k},
\end{aligned}$$

where k, indicates which class the examples in the estimate correspond to, C is the number of classes, $\mathbf{K}_{\Phi_{\text{dg}}} \in \mathbb{R}^{n_k \times n_k}$, $\mathbf{K}_{\Phi_{\text{spu}}} \in \mathbb{R}^{n_k \times n_k}$ are Gram matrices, $\mathbf{K}^{i,j}_{\Phi} = \kappa(\Phi_{\text{dg}}(X)_i, \Phi_{\text{dg}}(X)_j)$, $\mathbf{K}^{i,j}_{\Phi_{\text{spu}}} = \omega(\Phi_{\text{spu}}(X)_i, \Phi_{\text{spu}}(X)_j)$ with kernels κ, ω are radial basis functions, $\mathbf{H}_{n_k} = \mathbf{I}_{n_k} - \frac{1}{n_k^2} \mathbf{1}\mathbf{1}^\top$ is a centering matrix, $\mathbf{I}_{n_k}$ is the $n_k \times n_k$ dimensional identity matrix, $\mathbf{1}_{n_k}$ is the n_k-dimensional vector whose elements are all 1, and $^\top$ denotes the transpose. We condition on the label by taking only examples of each label and computing the empirical HSIC; then, we take the average.

Taken together, the full objective to be minimized is as follows:

$$\mathcal{L} = \frac{1}{E_{tr}} \sum_{e \in \mathcal{E}_{tr}} \Big[\mathcal{R}^e(\theta_c \circ \Phi_{\text{dg}}) + \mathcal{R}^e(\theta_e \circ \Phi) + \beta \mathcal{L}_{TCRI}(\Phi_{\text{dg}}, \Phi_{\text{spu}}) \Big],$$

where $\beta > 0$ is a hyperparameter and E_{tr} is the number of training domains. Figure 3 shows the full framework. We note that when $\beta = 0$, this loss reduces to ERM.

Note that while we minimize this objective with respect to $\Phi, \theta_c, \theta_1, \ldots, \theta_{E_{tr}}$, only the domain-general representation and its predictor, $\theta_c \cdot \Phi_{\text{dg}}$ are used for inference.

5 Experiments

We begin by evaluating with simulated data, i.e., with known ground truth mechanisms; we use Equation 5 to generate our simulated data, with domain parameter σ_{e_i}; code is provided in the supplemental materials.

$$\mathcal{SCM}(e_i) := \begin{cases} Z_{\text{dg}}^{(e_i)} \sim \mathcal{N}\left(0, \sigma_{e_i}^2\right) \\ y^{(e_i)} = Z_{\text{dg}}^{(e_i)} + \mathcal{N}\left(0, \sigma_y^2\right), \\ Z_{\text{spu}}^{(e_i)} = Y^{(e_i)} + \mathcal{N}\left(0, \sigma_{e_i}^2\right). \end{cases} \tag{5}$$

We observe 2 domains with parameters $\sigma_{e=0} = 0.1$, $\sigma^{e=1} = 0.2$ with $\sigma_y = 0.25$, 5000 samples, and linear feature extractors and predictors. We use partial covariance as our conditional independence penalty $\mathcal{L}_{TCRI}$. Table 2 shows the learned value of Φ_{dg}, where 'Oracle' indicates the true coefficients obtained by regressing Y on domain-general Z_{dg} directly. The ideal Φ_{dg} recovers Z_{dg} and puts zero weight on Z_{spu}.

Table 2: Continuous Simulated Results – Feature Extractor with a dummy predictor $\theta_c = 1.$, i.e., $\widehat{y} = x \cdot \Phi_{\text{dg}} \cdot w$, where $x \in \mathbb{R}^{N\times 2}$, $\Phi_{\text{dg}}, \Phi_{\text{spu}} \in \mathbb{R}^{2\times 1}$, $w \in \mathbb{R}$. Oracle indicates the coefficients achieved by regressing y on z_c directly.

Algorithm	$(\Phi_{\text{dg}})_0$ (i.e., Z_{dg} weight)	$(\Phi_{\text{dg}})_1$ (i.e., Z_{spu} weight)
ERM	0.29	0.71
IRM	0.28	0.71
TCRI	1.01	0.06
Oracle	1.04	0.00

Now, we evaluate the efficacy of our proposed objective on non-simulated datasets.

5.1 Semisynthetic and Real-World Datasets

Algorithms: We compare our method to baselines corresponding to DAG properties: Empirical Risk Minimization (**ERM**, [Vapnik, 1991]), Invariant Risk Minimization (**IRM** [Arjovsky et al., 2019]), Variance Risk Extrapolation (**V-REx**, [Krueger et al., 2021]), [Li et al., 2018a]), Group Distributionally Robust Optimization (**GroupDRO**), [Sagawa et al., 2019]), **Fish** [Shi et al., 2021], **CausIRL** [Chevalley et al., 2022], and Information Bottleneck methods (**IB_ERM**/**IB_IRM**, [Ahuja et al., 2021]). Additional baseline methods are provided in the Appendix A.

We evaluate our proposed method on the semisynthetic ColoredMNIST [Arjovsky et al., 2019] and real-world Terra Incognita dataset [Beery et al., 2018]. Given observed domains $\mathcal{E}_{tr} = \{e : 1, 2, \ldots, E_{tr}\}$, we train on $\mathcal{E}_{tr} \setminus e_i$ and evaluate the model on the unseen domain e_i, for each $e \in \mathcal{E}_{tr}$.

ColoredMNIST: The ColoredMNIST dataset [Arjovsky et al., 2019] is composed of 7000 ($2\times28\times28$, 1) images of a hand-written digit and binary-label pairs. There are three domains with different correlations between image color and label, i.e., the image color is spuriously related to the label by assigning a color to each of the two classes (0: digits 0-4, 1: digits 5-9). The color is then flipped with probabilities $\{0.1, 0.2, 0.9\}$ to create three domains, making the color-label relationship domain-specific because it changes across domains. There is also label flip noise of 0.25, so we expect that the best accuracy a domain-general model can achieve is 75%, while a non-domain general model can achieve higher. In this dataset, Z_{dg} corresponds to the original image, Z_{spu} the color, e the label-color correlation, Y the image label, and X the observed colored image. This DAG follows the generative process of Figure 2a [Arjovsky et al., 2019].

Spurrious PACS: Variables. X: images, Y: non-urban (elephant, giraffe, horse) vs. urban (dog, guitar, house, person). *Domains.* {{cartoon, art painting}, {art painting, cartoon}, {photo}} [Li et al., 2017]. The photo domain is the same as in the original dataset. In the {cartoon, art painting} domain, urban examples are selected from the original cartoon domain, while non-urban examples are selected from the original art painting domain. In the {art painting, cartoon} domain, urban examples are selected from the original art painting domain, while non-urban examples are selected from the original cartoon domain. This sampling encourages the model to use spurious correlations (domain-related information) to predict the labels; however, since these relationships are flipped between domains {{cartoon, art painting} and {art painting, cartoon}, these predictions will be wrong when generalized to other domains.

Terra Incognita: The Terra Incognita dataset contains subsets of the Caltech Camera Traps dataset [Beery et al., 2018] defined by [Gulrajani and Lopez-Paz, 2020]. There are four domains representing different locations {L100, L38, L43, L46} of cameras in the American Southwest. There are 9 species of wild animals {bird, bobcat, cat, coyote, dog, empty, opossum, rabbit, raccoon, squirrel} and a 'no-animal' class to be predicted. Like Ahuja et al. [2021], we classify this dataset as following the generative process in Figure 2c, the Fully Informative Invariant Features (FIIF) setting. Additional details on model architecture, training, and hyperparameters are detailed in Appendix 5.

Model Selection. The standard approach for model selection is a within-domain hold-out validation set. We find that model selection across hyperparameters using held-out training domain validation accuracy often returns non-domain-general models in the 'hard' cases. One advantage of our model is that we can do model selection based on the TCRI condition (conditional independence between the two representations) on held-out training domain validation examples in the 'hard' cases. In the easy case, we expect the empirical risk minimizer to be domain-general, so selecting the best-performing training-domain model is sound – we additionally do this for all baselines (see Appendix A.1 for further discussion). We find that, empirically, this heuristic works in the examples we study in this work. Nevertheless, model selection under distribution shift remains a significant bottleneck for domain generalization.

5.2 Results and Discussion

Worst-domain Accuracy. A critical implication of domain generality is stability – robustness in worst-domain performance up to domain difficulty. While average accuracy across domains provides some insight into an algorithm's ability to generalize to new domains, the average hides the variance of performance across domains. Average improvement can be increased while the worst-domain accuracy stays the same or decreases, leading to incorrect conclusions about domain generalization. Additionally, in real-world challenges such as algorithmic fairness where worst-group performance is considered, some metrics or fairness are analogous to achieving domain generalization Creager et al. [2021].

Results. TCRI achieves the highest average and worst-case accuracy across all baselines (Table 3). We find no method recovers the exact domain-general model's accuracy of 75%. However, TCRI achieves over 7% increase in both average accuracy and worst-case accuracy. Appendix A.2 shows transfer accuracies with cross-validation on held-out test domain examples (oracle) and TCRI again outperforms all baselines, achieving an average accuracy of 70.0% ± 0.4% and a worst-case accuracy of 65.7% ± 1.5, showing that regularizing for TCRI gives very close to optimal domain-general solutions.

Similarly, for the Spurious-PACS dataset, we observe that TCRI outperforms the baselines. TRCI achieves the highest average accuracy of $63.4\% \pm 0.2$ and worst-case accuracy of $62.3\% \pm 0.1$ with the next best, VREx, achieving 58.8 ± 1.0 and 33.8 ± 0.0, respectively. Additionally, for the Terra-Incognita dataset, TCRI achieves the highest average and worst-case accuracies of 49.2% ± 0.3% and 40.4% ± 1.6% with the next best, GroupDRO, achieving 47.8 ± 0.9 and 39.9 ± 0.7, respectively. Appendix A.2 shows transfer accuracies with cross-validation held-out target domain examples (oracle) where we observe that TCRI also obtains the highest average and worst-case accuracy for Spurrious-PACS and Terra Incognita.

Overall, regularizing for TCRI gives the most domain-general solutions compared to our baselines, achieving the highest worst-case accuracy on all benchmarks. Additionally, TCRI achieves the highest average accuracy on ColoredMNIST and Spurious-PAC and the second highest on Terra Incognita, where we expect the empirical risk minimizer to be domain-general.

Additional results are provided in the Appendix A.

The Effect of the Total Information Criterion. Without the TIC loss term, our proposed method is less effective. Table 5 shows that for Colored MNIST, the hardest 'hard' case we encounter, removing the TIC criteria, performs worse in average and worst case accuracy, dropping over 8% and 18, respectively.

Separation of Domain General and Domain Specific Features. In the case of Colored MNIST, we can reason about the extent of feature disentanglement from the accuracies achieved by the domain-general and domain-specific predictors. Table 4 shows how much each component of Φ, Φ_{dg} and Φ_{spu}, behaves as expected. For each domain, we observe that the domain-specific predictors' accuracies follow the same trend as the color-label correlation, indicating that they capture the color-label relationship. The domain-general predictor, however, does not follow such a trend, indicating that it is not using color as the predictor.

For example, when evaluating the domain-specific predictors from the +90% test domain experiment (row +90%) on held-out examples from the +80% training domain (column "DS Classifier on +80%"), we find that the +80% domain-specific predictor achieves an accuracy of nearly 79.9% – exactly what one would expect from a predictor that uses a color correlation with the same direction '+'. Conversely, the -90% predictor achieves an accuracy of 20.1%, exactly what one would expect from a predictor that uses a color correlation with the opposite direction '-'. The -90% domain has the opposite label-color pairing, so a color-based classifier will give the opposite label in any '+' domain.

Another advantage of this method, exemplified by Table 4, is that if one believes a particular domain is close to one of the training domains, one can opt to use the close domain's domain-specific predictor and leverage spurious information to improve performance.

On Benchmarking Domain Generalization. Previous work on benchmarking domain generalization showed that across standard benchmarks, the domain-

Table 3: $\mathcal{E}\backslash e_{test} \rightarrow e_{test}$ (model selection on held-out source domains validation set). The 'mean' column indicates the average generalization accuracy over all three domains as the e_{test} distinctly; the 'min' column indicates the worst generalization accuracy.

	ColoredMNIST		Spurious PACS		Terra Incognita	
Algorithm	**average**	**worst-case**	**average**	**worst-case**	**average**	**worst-case**
ERM	51.6 ± 0.1	10.0 ± 0.1	57.2 ± 0.7	31.2 ± 1.3	44.2 ± 1.8	35.1 ± 2.8
IRM	51.7 ± 0.1	9.9 ± 0.1	54.7 ± 0.8	30.3 ± 0.3	38.9 ± 3.7	32.6 ± 4.7
GroupDRO	52.0 ± 0.1	9.9 ± 0.1	58.5 ± 0.4	37.7 ± 0.7	47.8 ± 0.9	39.9 ± 0.7
VREx	51.7 ± 0.2	10.2 ± 0.0	58.8 ± 0.4	37.5 ± 1.1	45.1 ± 0.4	38.1 ± 1.3
FISH	51.7 ± 0.2	10.1 ± 0.0	57.8 ± 1.1	34.9 ± 2.4	47.2 ± 1.8	39.5 ± 2.3
Causal IRL	51.6 ± 0.2	10.2 ± 0.1	56.8 ± 0.8	34.4 ± 0.7	48.6 ± 0.3	38.4 ± 1.5
IB_ERM	51.5 ± 0.2	10.0 ± 0.1	56.3 ± 1.1	35.5 ± 0.4	46.0 ± 1.4	39.3 ± 1.1
IB_IRM	51.7 ± 0.0	9.9 ± 0.0	55.9 ± 1.2	33.8 ± 2.2	37.0 ± 2.8	29.6 ± 4.1
TCRI_HSIC	**59.6 ± 1.8**	**45.1 ± 6.7**	**63.4 ± 0.2**	**62.3 ± 0.2**	**49.2 ± 0.3**	**40.4 ± 1.6**

Table 4: Total Information Criterion: Domain General (DG) and Domain Specific (DS) Accuracies. The DG classifier is shared across all training domains, and the DS classifiers are trained on each domain. The first row indicates the domain from which the held-out examples are sampled, and the second indicates which domain-specific predictor is used. {+90%, +80%, -90%} indicate domains – $\{0.1, 0.2, 0.9\}$ digit label and color correlation, respectively.

	DG Classifier			**DS Classifier on +90**			**DS Classifier on +80**			**DS Classifier on -90**		
Test Domain	+90%	+80%	-90%	+90%	+80%	-90%	+90%	+80%	-90%	+90%	+80%	-90%
No DS clf.												
+90%	68.7	69.0	68.5	-	90.1	9.8	-	79.9	20.1	-	10.4	89.9
+80%	63.1	62.4	64.4	76.3	-	24.3	70.0	-	30.4	24.5	-	76.3
-90%	65.6	63.4	44.1	75.3	75.3	-	69.2	69.5	-	29.3	26.0	-

Table 5: TIC ablation for Colored MNIST.

Algorithm	**average**	**worst-case**
TCRI_HSIC (No TIC)	51.8 ± 5.9	27.7 ± 8.9
TCRI_HSIC	**59.6 ± 1.8**	**45.1 ± 6.7**

unaware empirical risk minimizer outperforms or achieves equivalent performance to the state-of-the-art domain generalization methods [Gulrajani and Lopez-Paz, 2020]. Additionally, Rosenfeld et al. [2022] gives results that show weak conditions that define regimes where the empirical risk minimizer across domains is optimal in both average and worst-case accuracy. Consequently, to accurately evaluate our work and baselines, we focus on settings where it is clear that (i) the empirical risk minimizer fails, (ii) spurious features, as we have defined them, do not generalize across the observed domains, and (iii) there is room for improvement via better domain-general predictions. We discuss this point further in the Appendix A.1.

Oracle Transfer Accuracies. While model selection is an integral part of the machine learning development cycle, it remains a non-trivial challenge when there is a distribution shift. While we have proposed a selection process tailored to our method that can be generalized to other methods with an assumed causal graph, we acknowledge that model selection under distribution shift is still an important open problem. Consequently, we disentangle this challenge from the learning problem and evaluate an algorithm's capacity to give domain-general solutions independently of model selection. We report experimental reports using held-out test-set examples for model selection in Appendix A Table 6. We find that our method, TCRI_HSIC, also outperforms baselines in this setting.

6 Conclusion and Future Work

We reduce the gap in learning domain-general predictors by leveraging conditional independence properties implied by generative processes to identify domain-general mechanisms. We do this without independent observations of domain-general and spurious mechanisms and show that our framework outperforms other state-of-the-art domain-generalization algorithms on real-world datasets in average and worst-case across domains. Future work includes further improvements to the framework to fully recover the strict set of domain-general mechanisms and model selection strategies that preserve desired domain-general properties.

Acknowledgements

OS was partially supported by the UIUC Beckman Institute Graduate Research Fellowship, NSF-NRT 1735252. This work is partially supported by the NSF III 2046795, IIS 1909577, CCF 1934986, NIH 1R01MH116226-01A, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc.

References

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

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

Martín Arjovsky, L. Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

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

Shai Ben-David, John Blitzer, K. Crammer, A. Kulesza, Fernando C Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2009.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR, 2016.

Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Rémi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *ArXiv*, abs/2003.04475, 2020.

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

A. Gretton, K. Fukumizu, C. Teo, Le Song, B. Schölkopf, and Alex Smola. A kernel statistical test of independence. In *NIPS*, 2007.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *CoRR*, abs/2007.01434, 2020. URL https://arxiv.org/abs/2007.01434.

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

Christopher Hitchcock and Miklós Rédei. Reichenbach's Common Cause Principle. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, 19, 2006.

Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint arXiv:2206.07837*, 2022.

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

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1882–1886. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/kpotufe18a.html.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai

Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

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

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018b. doi: 10.1109/CVPR.2018.00566.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018c.

Zachary Chase Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. *ArXiv*, abs/1802.03916, 2018.

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

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. *ArXiv*, abs/1502.02791, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 739–766. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/makar22a.html.

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

J. Pearl. Causal inference. In *NIPS Causality: Objectives and Assessment*, 2010.

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)*, pages 947–1012, 2016.

Miklós Rédei. *Reichenbach's Common Cause Principle and Quantum Correlations*, pages 259–270. Springer Netherlands, Dordrecht, 2002. ISBN 978-94-010-0385-8. doi: 10.1007/978-94-010-0385-8_17. URL https://doi.org/10.1007/978-94-010-0385-8_17.

Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229, 2021.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2641–2657. PMLR, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

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

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems*, 20, 2007.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *NIPS*, volume 91, pages 831–840, 1991.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *ICML*, 2022.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114, 2004.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34, 2021.

H. Zhao, Rémi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *ICML*, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No] We are planning to release the code upon publication.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplements

A Additional Results and Discussion

A.1 On Benchmarking Domain Generalization

Table 6: Oracle (model selection on held-out target domain validation set) $\mathcal{E}\backslash e_{test} \rightarrow e_{test}$. The 'mean' column indicates the average generalization accuracy over all three domains as the e_{test} distinctly; the 'min' column indicates the worst generalization accuracy.

	ColoredMNIST		Spurious PACS		Terra Incognita	
Algorithm	**average**	**worst-case**	**average**	**worst-case**	**average**	**worst-case**
ERM	57.8 ± 0.2	38.4 ± 1.4	59.2 ± 1.3	38.4 ± 1.4	**52.9 ± 0.8**	42.0 ± 0.6
IRM	68.9 ± 1.6	62.0 ± 4.9	67.5 ± 5.8	53.9 ± 6.6	42.6 ± 4.0	42.7 ± 1.2
GroupDRO	61.1 ± 1.3	37.6 ± 3.6	61.8 ± 1.8	40.0 ± 1.6	50.7 ± 1.0	42.7 ± 1.2
VREx	68.0 ± 2.5	59.4 ± 7.3	62.8 ± 2.4	38.7 ± 0.9	43.2 ± 2.0	34.9 ± 4.2
IB_ERM	65.0 ± 0.1	50.6 ± 0.3	67.3 ± 3.7	53.1 ± 8.0	49.0 ± 0.3	39.9 ± 0.8
IB_IRM	68.4 ± 1.0	58.5 ± 2.8	69.0 ± 1.3	**62.3 ± 0.3**	32.8 ± 6.6	20.4 ± 7.5
TCRI_HSIC (TCRI)	**70.4 ± 0.4**	**65.7 ± 1.5**	**69.5 ± 1.1**	**62.3 ± 0.2**	51.2 ± 0.1	**43.0 ± 0.4**

Oracle Transfer Accuracies. While model selection is an integral part of the machine learning development cycle, it remains a non-trivial challenge when there is a distribution shift. While we have proposed a selection process tailored to our method that can be generalized to other methods with an assumed causal graph, we acknowledge that model selection under distribution shift is still an important open problem. Consequently, we disentangle this challenge from the learning problem and evaluate an algorithm's capacity to give domain-general solutions independently of model selection. We report experimental reports using held-out test-set examples for model selection in Appendix A Table 6.

In this case, we find that there is indeed a separation between ERM and some domain-generalization algorithms, suggesting that model selection might be the bottleneck for learning domain-general predictors. Nevertheless, we still find that our method, TCRI_HSIC, also outperforms baselines in this setting.

Challenges of Benchmarking Domaing Generalization. We show some results below that illustrate the challenge of accurately evaluating the efficacy of an algorithm for domain generalization. We first note that we expect ERM (naive) to perform poorly in domain generalization tasks, certainly so when we observe worst-case shifts at test time. However, like other works [Gulrajani and Lopez-Paz, 2020], we observe that ERM performs as well as other baselines during transfer on various benchmark datasets. Previous theoretical results [Rosenfeld et al., 2022] suggest that this observation is indicative of properties of the benchmark domains that may be sufficient for ERM to give domain-general solutions - specifically that the distribution (and equivalently the loss) of the target domain can be written as a convex combination of the those in the source domains.

To further investigate this, we develop additional experiments motivated by the ColoredMNIST [Arjovsky et al., 2019] – since its generative process is well understood. We note that in the +90%, +80%, and -90% domains of ColoredMNIST, the -90% domain has the opposite relationship between the spurious correlation and the label, so the use of spurious correlations from {+90%, +80%} generalizes catastrophically to the -90% domain. In this setting, the baseline algorithms we present, including ERM, achieve poor accuracy in the -90% domain while maintaining high accuracy in the +90% and +80% domains. Consequently, we investigate two settings, *setting a*: observe {+90%, +80%, +70%, -90%} domains and *setting b*: observe {+90%, +80%, -80%, -90%} domains – we focus on generalizing to the -90% domain. In *setting a*, we add another domain with the majority direction in the relationship between spurious correlation and labels. In *setting b*, we add another domain with the minority

direction. Note that in *setting a*, the closest domain to -90% that can be generated with a convex combination of the other domains still has a '+' correlation between the color and label. In *setting b*, however, one can generate a domain with a '-' correlation between color and label with a convex combination of the other domains. Thus, we expect the empirical risk minimizer to give domain-general solutions in setting b but not in setting a.

We use Oracle model selection (held-out target data) to remove the effect of model selection for all methods in the results. We find that in setting a, where we add a domain (+70%), we observe that the generalization accuracy to the -90% domain is still very different from the other domains (Table 7).

Table 7: Colored MNIST *setting a*. Columns {+90%, +80%, +70%, -90%} indicate domains – $\{0.1, 0.2, 0.3, 0.9\}$ digit label and color correlation, respectively. We report domain accuracies over 3 trials each. We use the oracle selection method – held out target data. $\mathcal{E}\backslash e_{test} \rightarrow e_{test}$.

Algorithm	**+90%**	**+80%**	**+70%**	**-90%**
ERM	72.8 ± 0.3	74.7 ± 0.3	73.3 ± 0.1	16.3 ± 1.5
IRM	49.0 ± 0.1	54.2 ± 2.0	50.3 ± 0.3	43.8 ± 2.8
GroupDRO	71.0 ± 0.6	72.2 ± 0.3	70.7 ± 0.9	36.4 ± 4.2
VREx	74.1 ± 1.3	72.6 ± 0.5	72.1 ± 0.5	19.5 ± 5.5
TCRI (HSIC)	72.1 ± 1.5	73.6 ± 0.4	72.6 ± 0.4	49.9 ± 0.3

However, in setting b, where we add a domain (-80%), we observe that the generalization accuracy to the -90% domain is on par with the other domains (Table 8).

Table 8: Colored MNIST *setting b*. Columns {+90%, +80%, -80%, -90%} indicate domains – $\{0.1, 0.2, 0.8, 0.9\}$ digit label and color correlation, respectively. We report the average domain accuracies over 3 trials each. We use the oracle selection method – held out target data. $\mathcal{E}\backslash e_{test} \rightarrow e_{test}$.

Algorithm	**+90%**	**+80%**	**-80%**	**-90%**
ERM	58.4 ± 1.3	67.0 ± 0.5	64.2 ± 2.0	52.6 ± 3.2
IRM	56.7 ± 3.3	56.6 ± 2.8	51.6 ± 0.7	51.7 ± 0.7
GroupDRO	69.7 ± 0.8	71.7 ± 0.3	72.0 ± 0.2	71.4 ± 1.9
VREx	67.4 ± 1.9	70.4 ± 0.1	71.2 ± 0.2	59.4 ± 4.3
TCRI (HSIC)	62.2 ± 4.4	70.0 ± 1.3	67.9 ± 1.4	65.4 ± 2.8

This illustrates the challenge of accurately evaluating an algorithm's ability to give domain-general predictions. We note that it is generally difficult to distinguish between *setting a* and *setting b*. The primary signature we see is some consistency between the empirical risk minimizer and the other baselines. Gulrajani and Lopez-Paz [2020] observe a similar trend for standard benchmarks for domain generalization. Hence, we focus our empirical evaluations in this work on settings where we know that the ERM solution fails by design.

A.2 ColoredMNIST

ColoredMNIST: The ColoredMNIST dataset [Arjovsky et al., 2019] is composed of 7000 $(2 \times 28 \times 28, 1)$ images of a hand-written digit and binary-label pairs. There are three domains with different correlations between image color and label, i.e., the image color is spuriously related to the label by assigning a color to each of the two classes (0: digits 0-4, 1: digits 5-9). The color is then flipped with probabilities $\{0.1, 0.2, 0.9\}$ to create three domains, making the color-label relationship domain-specific because it changes across domains. There is also label flip noise of 0.25, so we expect that the best accuracy a domain-general model can achieve is 75%, while a non-domain general model can achieve higher. In this dataset, Z_{dg} corresponds to the original image, Z_{spu} the color, e the label-color correlation, Y the image label, and X the observed colored image. This DAG follows the generative process of Figure 2a

We use MNIST-ConvNet Gulrajani and Lopez-Paz [2020] backbones for the MNIST datasets (Table 10). Both Φ_{dg} and Φ_{spu} are linear layers of size 128×128 that are appended to the backbone. The predictors (classification hyperplanes) θ_c, $\{\theta_1, \theta_2\}$ are also parameterized to be linear and appended to the Φ_{dg} and Φ, respectively.

Table 9: Colored MNIST Hyperparameters. Additional hyperparameters provided in https://github.com/olawalesalaudeen/tcri.

Algorithm	Hyperparameter	Default	Random Distribution
All	Learning Rate	1^{-3}	$10^{\text{Uniform}(-4.5,-2.5)}$
	Batch Size	64	$2^{\text{Uniform}(3,9)}$
TCRI β	penalty weight	100	$10^{\text{Uniform}}(-1,5)$
	annealing steps	500	$10^{\text{Uniform}}(2.5,5)$

Table 10: MNIST ConvNet architecture. All convolutions use 3×3 kernels and "same" padding.

#	**Layer**
1	Conv2D (in=d, out=64)
2	ReLU
3	GroupNorm (groups=8)
4	Conv2D (in=64, out=128, stride=2)
5	ReLU
6	GroupNorm (groups=8)
7	Conv2D (in=128, out=128)
8	ReLU
9	GroupNorm (groups=8)
10	Conv2D (in=128, out=128)
11	ReLU
12	GroupNorm (8 groups)
13	Global average-pooling

We do a random search to select hyperparameters using the same scheme as Gulrajani and Lopez-Paz [2020] (https://github.com/facebookresearch/DomainBed). We select 25 hyperparameters with 5 random restarts each to generate error bars.

We show transfer accuracies with both source and target domain validation for model selection in Tables 11-12. We find that TCRI outperforms all baselines in average and worst-case accuracy.

Table 11: Colored MNIST Transfer Accuracy – model selection on held-out source validation set. Columns {+90%, +80%, -90%} indicate domains – $\{0.1, 0.2, 0.9\}$ digit label and color correlation, respectively. $\mathcal{E}\backslash e_{test} \rightarrow e_{test}$.

	Domains			Domain Accuracy Statistics		
Algorithm	**+90%**	**+80%**	**-90%**	**Avg**	**Std**	**Min**
ERM	71.6 ± 0.3	73.1 ± 0.1	10.0 ± 0.1	51.6 ± 0.1	29.4 ± 0.1	10.0 ± 0.1
IRM	72.1 ± 0.1	73.0 ± 0.3	9.9 ± 0.1	51.7 ± 0.1	29.5 ± 0.1	9.9 ± 0.1
GroupDRO	72.6 ± 0.2	73.4 ± 0.2	9.9 ± 0.1	52.0 ± 0.1	29.8 ± 0.1	9.9 ± 0.1
VREx	72.2 ± 0.2	72.7 ± 0.3	10.2 ± 0.0	51.7 ± 0.2	29.3 ± 0.1	10.2 ± 0.0
IB_ERM	71.0 ± 0.4	73.4 ± 0.3	10.0 ± 0.1	51.5 ± 0.2	29.4 ± 0.1	10.0 ± 0.1
IB_IRM	71.7 ± 0.2	73.4 ± 0.1	9.9 ± 0.0	51.7 ± 0.0	29.5 ± 0.0	9.9 ± 0.0
TCRI_HSIC	67.2 ± 2.3	65.6 ± 3.4	45.9 ± 6.9	**59.6 ± 1.8**	**11.4 ± 3.3**	**45.1 ± 6.7**

A.3 Spurrious PACS

Spurious–PACS. Variables. X: images, Y: non-urban (elephant, giraffe, horse) vs. urban (dog, guitar, house, person). *Domains.* {{cartoon, art painting}, {art painting, cartoon}, {photo}} [Li et al., 2017]. The photo domain is the same as in the original dataset. In the {cartoon, art painting} domain, urban examples are selected from the original cartoon domain, while non-urban examples are selected from the original art painting domain. In the {art painting, cartoon} domain, urban examples are selected from the original art painting domain, while non-urban examples are selected from the original cartoon domain. This sampling encourages the model to use spurious correlations (domain-related information) to predict the labels; however, since these relationships are

Table 12: Colored MNIST Transfer Accuracy – model selection on held-out target validation set accuracy. Columns {+90%, +80%, -90%} indicate domains – $\{0.1, 0.2, 0.9\}$ digit label and color correlation, respectively. $\mathcal{E}\backslash e_{test} \to e_{test}$.

	Domains			Domain Accuracy Statistics		
Algorithm	**+90%**	**+80%**	**-90%**	**Avg**	**Std**	**Min**
ERM	71.7 ± 0.4	73.4 ± 0.1	28.3 ± 0.2	57.8 ± 0.2	20.9 ± 0.2	28.3 ± 0.2
IRM	72.3 ± 0.3	72.4 ± 0.1	62.0 ± 4.9	68.9 ± 1.6	5.3 ± 2.0	61.5 ± 4.5
GroupDRO	73.5 ± 0.4	72.4 ± 0.1	37.6 ± 3.6	61.1 ± 1.3	16.7 ± 1.7	37.6 ± 3.6
VREx	72.0 ± 0.1	72.6 ± 0.4	59.4 ± 7.3	68.0 ± 2.5	7.6 ± 2.2	57.7 ± 6.0
IB_ERM	71.3 ± 0.2	73.2 ± 0.2	50.6 ± 0.3	65.0 ± 0.1	10.2 ± 0.2	50.6 ± 0.3
IB_IRM	74.1 ± 0.9	72.7 ± 0.4	58.5 ± 2.8	68.4 ± 1.0	7.1 ± 1.2	58.5 ± 2.8
TCRI_HSIC	72.8 ± 0.4	72.7 ± 0.2	65.7 ± 1.5	**70.4 ± 0.4**	**3.4 ± 0.7**	**65.7 ± 1.5**

flipped between domains {{cartoon, art painting} and {art painting, cartoon}, these predictions will be wrong when generalized to other domains.

Table 13: Spurrious PACS Hyperparameters. Additional hyperparameters provided in https://github.com/olawalesalaudeen/tcri.

Algorithm	Hyperparameter	Default	Range
All	Learning Rate	1^{-3}	$10^{\text{Uniform}(-4.5,-2.5)}$
	Batch Size	64	$2^{\text{Uniform}(3,9)}$
TCRI β	penalty weight	100	$10^{\text{Uniform}}(-1, 5)$
	annealing steps	500	$10^{\text{Uniform}}(2.5, 5)$

We use a ResNet-50 backbone [He et al., 2016]. Φ_{dg} and Φ_{spu} are linear layers of size 2048×2048 that are appended to the backbone. The predictors (classification hyperplanes) $\theta_c, \{\theta_1, \theta_2, \theta_3\}$ are linear and appended to Φ_{dg} and Φ layers, respectively.

Hyperparameters: We do a random search to select hyperparameters using the same scheme as Gulrajani and Lopez-Paz [2020] (https://github.com/facebookresearch/DomainBed). We select 5 hyperparameters with 3 random restarts each to generate error bars.

We show transfer accuracies with both source and target domain validation for model selection in Tables 14-15. We find that TCRI outperforms all baselines in average and worst-case accuracy.

Table 14: Spurious–PACS Transfer Accuracy – model selection on held-out source validation set. $\mathcal{E}\backslash e_{test} \to e_{test}$.

	Domains			Domain Accuracy Statistics		
Algorithm	**C x A**	**A x C**	**P**	**mean**	**std**	**min**
ERM	31.2 ± 1.3	42.8 ± 0.7	97.6 ± 0.2	57.2 ± 0.7	29.0 ± 0.4	31.2 ± 1.3
IRM	30.3 ± 0.3	39.0 ± 1.3	94.9 ± 1.4	54.7 ± 0.8	28.6 ± 0.8	30.3 ± 0.3
GroupDRO	37.7 ± 0.7	42.1 ± 1.6	95.7 ± 0.5	58.5 ± 0.4	26.4 ± 0.3	37.7 ± 0.
VREx	37.5 ± 1.1	43.0 ± 0.5	95.7 ± 1.5	58.8 ± 0.4	26.2 ± 1.0	37.5 ± 1.1
IB_ERM	35.5 ± 0.4	48.6 ± 3.3	84.8 ± 0.6	56.3 ± 1.1	20.8 ± 0.6	35.5 ± 0.4
IB_IRM	33.8 ± 2.2	38.8 ± 3.0	95.1 ± 1.5	55.9 ± 1.2	27.8 ± 1.5	33.8 ± 0.4
TCRI (HSIC)	62.8 ± 0.1	62.3 ± 0.2	65.0 ± 0.4	**63.4 ± 0.2**	**1.2± 0.2**	**62.3 ± 0.2**

Table 15: Oracle Spurious–PACS Transfer Accuracy – model selection on held-out target validation set. $\mathcal{E}\backslash e_{test} \rightarrow e_{test}$.

	Domains		Domain Accuracy Statistics			
Algorithm	**C x A**	**A x C**	**P**	**mean**	**std**	**min**
ERM	38.4 ± 1.4	43.4 ± 1.9	95.9 ± 0.6	59.2	26.0	38.4
IRM	62.8 ± 0.1	53.9 ± 6.6	85.8 ± 8.2	67.5	13.4	53.9
GroupDRO	40.0 ± 1.6	49.7 ± 2.9	95.7 ± 0.6	61.8	24.3	40.0
VREx	55.8 ± 5.5	38.7 ± 0.9	93.8 ± 0.8	62.8	23.0	38.7
IB_ERM	53.1 ± 8.0	55.4 ± 5.7	93.5 ± 1.8	67.3	18.5	53.1
IB_IRM	62.8 ± 0.1	62.3 ± 0.3	81.8 ± 7.0	69.0	**9.1**	**62.3**
TCRI (HSIC)	64.0 ± 0.7	62.3 ± 0.2	82.4 ± 5.7	**69.5**	**9.1**	**62.3**

A.4 Terra Incognita

The Terra Incognita dataset contains subsets of the Caltech Camera Traps dataset [Beery et al., 2018] defined by [Gulrajani and Lopez-Paz, 2020]. Four domains represent different locations {L100, L38, L43, L46} of cameras in the American Southwest. There are 10 different species of wild animals {bird, bobcat, cat, coyote, dog, empty, opossum, rabbit, raccoon, squirrel} (classes) to be predicted. Like Ahuja et al. [2021], we classify this dataset as following the generative process in Figure 2c, the Fully Informative Invariant Features (FIIF) setting.

Table 16: Terra Incognita Hyperparameters. Additional hyperparameters provided in https://github.com/olawalesalaudeen/tcri.

Algorithm	Hyperparameter	Default	Range
All	Learning Rate	1^{-3}	$10^{\text{Uniform}(-4.5,-2.5)}$
	Batch Size	64	$2^{\text{Uniform}(3,9)}$
TCRI β	penalty weight	100	$10^{\text{Uniform}}(-1,5)$
	annealing steps	500	$10^{\text{Uniform}}(0,4)$

We use a ResNet-50 backbone [He et al., 2016]. Φ_{dg} and Φ_{spu} are linear layers of size 2048×2048 that are appended to the backbone. The predictors (classification hyperplanes) $\theta_c, \{\theta_1, \theta_2, \theta_3, \theta_4\}$ are linear and appended to Φ_{dg} and Φ layers, respectively.

Hyperparameters: We do a random search to select hyperparameters using the same scheme as Gulrajani and Lopez-Paz [2020] (https://github.com/facebookresearch/DomainBed). We select 5 hyperparameters with 3 random restarts each to generate error bars.

We show transfer accuracies with both source and target domain validation for model selection in Tables 17-18. We find that TCRI outperforms all baselines except ERM on average and outperforms all baselines in worst-case accuracy.

Table 17: Terra Incognita Transfer Accuracy – model selection on held-out source validation set. $\mathcal{E}\backslash e_{test} \rightarrow e_{test}$.

	Domains				Domain Accuracy Statistics		
Algorithm	**L100**	**L38**	**L43**	**L46**	**Avg**	**Std**	**Min**
ERM	43.6 ± 3.9	45.2 ± 0.6	53.0 ± 1.2	35.1 ± 2.8	44.2 ± 1.8	6.8 ± 1.0	35.1 ± 2.8
IRM	43.9 ± 3.3	35.7 ± 4.0	37.7 ± 7.8	38.3 ± 2.4	38.9 ± 3.7	**5.4 ± 1.8**	32.6 ± 4.7
GroupDRO	53.8 ± 4.6	40.5 ± 0.7	55.3 ± 1.5	41.8 ± 1.1	47.8 ± 0.9	7.7 ± 0.9	39.9 ± 0.7
VREx	48.8 ± 2.0	38.1 ± 1.3	54.4 ± 0.6	39.0 ± 1.4	45.1 ± 0.4	7.0 ± 0.9	38.1 ± 1.3
IB_ERM	46.1 ± 4.5	40.7 ± 0.7	55.2 ± 0.8	42.2 ± 1.1	46.0 ± 1.4	6.4 ± 0.8	39.3 ± 1.1
IB_IRM	39.7 ± 7.3	40.8 ± 2.3	34.7 ± 4.3	32.9 ± 2.6	37.0 ± 2.8	6.7 ± 1.3	29.6 ± 4.1
TCRI_HSIC	54.6 ± 2.4	48.6 ± 2.0	53.2 ± 1.0	40.4 ± 1.6	**49.2 ± 0.3**	6.1 ± 1.1	**40.4 ± 1.6**

Table 18: Terra Incognita Transfer Accuracy – model selection on held-out target validation set. $\mathcal{E}\backslash e_{test} \to e_{test}$.

	Domains				Domain Accuracy Statistics		
Algorithm	**L100**	**L38**	**L43**	**L46**	**Avg**	**Std**	**Min**
ERM	58.5 ± 1.8	52.0 ± 1.3	59.2 ± 0.2	42.0 ± 0.6	**52.9 ± 0.8**	7.0 ± 0.5	42.0 ± 0.6
IRM	53.0 ± 0.9	48.0 ± 1.8	36.3 ± 9.6	33.2 ± 3.9	42.6 ± 4.0	9.6 ± 1.7	30.8 ± 5.4
GroupDRO	56.2 ± 3.0	45.2 ± 2.3	58.0 ± 0.2	43.3 ± 0.7	50.7 ± 1.0	6.9 ± 0.9	42.7 ± 1.2
VREx	43.2 ± 1.5	49.3 ± 1.2	41.5 ± 7.8	38.9 ± 1.1	43.2 ± 2.0	6.5 ± 1.8	34.9 ± 4.2
IB_ERM	55.6 ± 1.7	47.2 ± 1.1	53.4 ± 0.7	39.9 ± 0.8	49.0 ± 0.3	6.4 ± 0.5	39.9 ± 0.8
IB_IRM	40.2 ± 8.2	31.9 ± 11.8	29.4 ± 4.4	29.7 ± 3.8	32.8 ± 6.6	8.2 ± 1.0	20.4 ± 7.5
TCRI_HSIC	57.7 ± 1.8	50.1 ± 1.8	54.1 ± 0.6	43.0 ± 0.4	51.2 ± 0.1	**5.8 ± 0.7**	**43.0 ± 0.4**

B DAGs

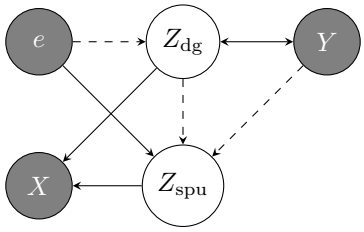


Figure 4: Partial Ancestral Graph (PAG). Dashed edges indicate that the edge may or may not exist. The combination of $Y \to Z_{\text{dg}} \to Z_{\text{spu}}$, and $Y \to Z_{\text{dg}}$, $e \to Z_{\text{dg}}$ is not allowed.

B.1 On Valid DAGS:

We consider other edges that could be introduced to Figure 4 where $Z_{\text{dg}} \not\perp\!\!\!\perp Z_{\text{spu}} \,|\, Y, e$, $Z_{\text{spu}} \not\perp\!\!\!\perp Y \,|\, Z_{\text{dg}}$, or are not included in Figure 5. and show that these edges either make the problem intractable or require new assumptions about the generative process – note we do not discuss edges that induce a cycle, thus, are invalid.

(i) $e - Y$: we do have a direct edge in either direction e between Y otherwise, Y is always dependent on e and the problem becomes intractable.

(ii) $e - X$: we do have a direct edge from $e - X$ without making additional parametric assumptions about the role of e in $\Gamma(Z_{\text{dg}}, Z_{\text{spu}}, e)$.

(iii) $Z_{\text{spu}} \to Y$: we do have both $Z_{\text{dg}} \to Y$ and $Z_{\text{spu}} \to Y$, since then, both mechanisms are domain general. WLOG, we let Z_{spu} denote the features that never have domain-general mechanisms to Y.

(iv) $Y \to Z_{\text{dg}} \to Z_{\text{spu}}$ and $Y \to Z_{\text{dg}} \leftarrow e$: conditioning on Z_{dg} and/or Z_{spu} make Y dependent on e, so Y is always dependent on e and the problem becomes intractable.

Table 19: Generative Processes and Sufficient Conditions for Domain-Generality

	Graphs in Figure 5		
	(a)	(b)	(c)
$Z_{\text{dg}} \perp\!\!\!\perp Z_{\text{spu}} \,|\, \{Y, e\}$	✓	✓	✗
Identifying Z_{dg} is necessary	✓	✓	✗

B.2 Fully Informative Invariant Features

We briefly summarize Ahuja et al. [2021]'s results on minimax-optimality of Empirical Risk Minimization in the Fullly Informative Invariant Features setting (their Lemma 4). First, we informally state their assumptions.

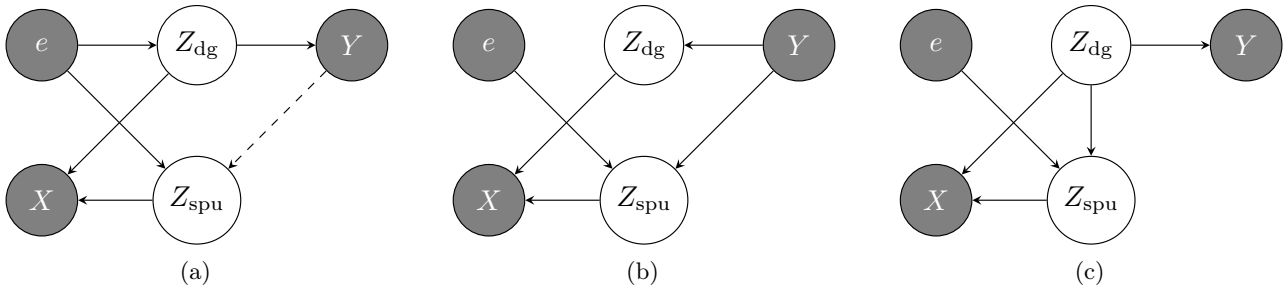


Figure 5: **Generative Processes.** Graphical model depicting the structure of our data-generating process - shaded nodes indicate observed variables. X represents the observed features, Y represents observed targets, and e represents domain influences. There is an explicit separation of domain-general Z_{dg} and domain-specific Z_{spu} features combined to generate observed X. Dashed edges indicate the possibility of an edge.

Table 20: Generative Processes and Sufficient Algorithms

	Graphs in Figure 5		
	(a)	(b)	(c)
Solved by ERM	✗	✗	✓
Solved by TCRI	✓	✓	✓

Assumption 2: Linear structural equation model.

Assumption 3-4: Bounded Features.

Assumption 8: w_{dg} partitions $\mathcal{Z}$ up to noise η_Y.

These assumptions are implied by our Assumption 4.1.

B.2.1 Proof Sufficiency of ERM [Ahuja et al., 2021]

If Assumptions 2, 4, and 8 hold, then there exists a classifier that puts a non-zero weight on the spurious feature and continues to be Bayes optimal in all the training environments.

Proof. Choose an arbitrary non-zero vector and derive a bound on the margin of (w_{dg}, γ), where w_{dg} is the true (optimal) linear predictor of Y from Z_{dg}. Recall domain-general and domain-specific features $z_{\text{dg}} \in \mathcal{Z}_{\text{dg}}$, $z_{\text{spu}} \in \mathcal{Z}_{\text{spu}}$, respectively. Let $y^* = \text{sign}(w_{\text{dg}} \cdot z_{\text{dg}})$. The margin of $(w_{\text{dg}}, \gamma))$ at point $(z_{\text{dg}}, z_{\text{spu}})$ with respect to y^* is defined as:

$$y^*(w_{\text{dg}} \cdot z_{\text{dg}}) + y^*(\gamma \cdot z_{\text{spu}}).$$

Using Cauchy-Schwartz inequality, we get

$$|y^*(\gamma \cdot z_{\text{spu}})| = |\gamma \cdot z_{\text{spu}}| \leq \|\|\gamma\| z_{\text{spu}}\|.$$

Since Z_{spu} is bounded, one can set γ sufficiently small enough to control $y^*(\gamma \cdot Z_{\text{spu}})$. If $\|\gamma\| \leq \frac{c}{2z^{\text{sup}}}$, then $|\gamma \cdot z_{\text{spu}}| \leq \frac{c}{2}$, where z^{sup} satisfies that $\|z\| \leq z^{\text{sup}} \forall z \in \mathcal{Z}_{\text{spu}}$. From Assumption 8, $\exists\, c > 0$ s.t.,

$$y^*(w_{\text{dg}} \cdot z_{\text{dg}}) \geq c.$$

Using $|\gamma \cdot z_{\text{spu}}| \leq \frac{c}{2}$, the margin becomes

$$y^*(w_{\text{dg}} \cdot z_{\text{dg}}) + y^*(\gamma \cdot z_{\text{spu}}) \geq c - |\gamma \cdot z_{\text{spu}}| \geq \frac{c}{2}.$$

From the above equation, it follows that $\text{sign}\left((w_{\text{dg}}, \gamma) \cdot (z_{\text{dg}}, z_{\text{spu}})\right) = \text{sign}\left((w_{\text{dg}}, 0) \cdot (z_{\text{dg}}, z_{\text{spu}})\right) \forall z_{\text{dg}} \in \mathcal{Z}_{\text{dg}}$, $z_{\text{spu}} \in \mathcal{Z}_{\text{spu}}$.

Now, this condition is used to compute the error of a spurious classifier, i.e., based on $(, \gamma)$. Define $g_{\text{spu}} = I \circ (w_{\text{dg}}, \gamma) \circ \Gamma^{-1}$, where $I(\cdot)$ is an indicator function that returns 1 if its input is ≥ 0. The error achieved by g_{spu} is

$$\begin{aligned} R^e(g_{\text{spu}}) &= \mathbb{E}\big[Y^e \oplus I((w_{\text{dg}}, \gamma) \cdot (z_{\text{dg}}, z_{\text{spu}})\big] \\ &= \mathbb{E}\Big[I\big((w_{\text{dg}}, 0) \cdot (z_{\text{dg}}, z_{\text{spu}})\big) \oplus \eta_y \oplus I\big((w_{\text{dg}}, \gamma) \cdot (z_{\text{dg}}, z_{\text{spu}})\big)\Big] \\ &= \mathbb{E}[\eta_y]. \end{aligned}$$

The error achieved by g_{spu} is then due to the noise in observed Y and is, therefore, optimal in all domains. □

It follows from above that since g_{spu} is Bayes optimal in every domain, it is also the empirical risk minimizer (ERM) as it minimizes the sum of risks across training domains.

C Proof of Proposition 4.3

Assume that $\Phi_{\text{dg}}(X)$ and $\Phi_{\text{spu}}(X)$ are correlated with Y. Given Assumptions 4.1-4.2 and a representation $\Phi = \Phi_{\text{dg}} \oplus \Phi_{\text{spu}}$ that satisfies TIC, $\Phi_{\text{dg}}(X) = Z_{\text{dg}} \iff \Phi$ satisfies TCRI. (see Appendix

Proof. 'only if'. Assume that $\Phi_{\text{dg}}(X) = Z_{\text{dg}}$. By the Total Information Criterion, we have that $\Phi_{\text{spu}}(X) = Z_{\text{spu}}$. We observe the following paths from Z_{dg} to Z_{spu}: (i) $Z_{\text{dg}} \to Y \to Z_{\text{spu}}$, (ii) $Z_{\text{dg}} \leftarrow e \to Z_{\text{spu}}$, and (iii) $Z_{\text{dg}} \to X \to Z_{\text{spu}}$. Conditioning on Y, e blocks both paths (i) and path (ii); path (iii) contains a collider (Z_{dg} and Z_{spu} are common causes of X), so this path is blocked when X is not in the conditioning set. So, $Z_{\text{spu}} \perp\!\!\!\perp Z_{\text{dg}} \,|\, Y, e$ and therefore $\Phi_{\text{dg}}(X) \perp\!\!\!\perp \Phi_{\text{spu}}(X) \,|\, Y, e$, which completes this direction.

'if'. Assume that Φ satisfies TCRI. We proceed by contradiction. Let $\Phi = [\Phi_{\text{dg}}; \Phi_{\text{spu}}]$. We consider the following scenario for $\Phi_{\text{dg}} \neq Z_{\text{dg}}$.

Scenario 1 (causal aggregation): Assume that $\Phi_{\text{dg}}(X) \subset Z_{\text{dg}}$. From TIC, we have that $Z^\dagger_{\text{dg}} \subset \Phi_{\text{spu}}(X)$, where $Z^\dagger_{\text{dg}} \subset Z_{\text{dg}}$ is the subset of Z_{dg} not captured by Φ_{dg}. Since $\Phi_{\text{dg}}(X)$ and $Z^\dagger_{\text{dg}}$ are colliders on Y, given both are subsets of Z_{dg}, $\Phi_{\text{dg}}(X) \not\perp\!\!\!\perp \Phi_{\text{spu}}(X) | Y, e$, violating TCRI and giving a contradiction. So, $Z_{\text{dg}} \subset \Phi(X)$

Scenario 2 (anticausal exclusion): Assume that $\Phi_{\text{dg}}(X) \subset Z_{\text{spu}}$. From TIC, we have that $Z^\dagger_{\text{spu}} \subset \Phi_{\text{spu}}(X)$, where $Z^\dagger_{\text{spu}} \subset Z_{\text{spu}}$ is the subset of Z_{spu} not captured by Φ_{dg}. From Assumption 4.2 (faithfulness), we have that $\Phi_{\text{dg}}(X) \not\perp\!\!\!\perp \Phi_{\text{spu}}(X) | Y, e$, violating TCRI and giving a contradiction. So, $Z_{\text{spu}} \not\subset \Phi_{\text{dg}}(X)$.

Combining scenarios 1-2, it follows that $\Phi_{\text{dg}}(X) = Z_{\text{dg}}$. □