Diagnosing and Rectifying Fake OOD Invariance: A Restructured Causal Approach

Ziliang Chen[1,2], Yongsen Zheng[3], Zhao-Rong Lai[1], Quanlong Guan[1*], Liang Lin[3]

[1]Jinan University
[2]Pazhou Lab
[3]Sun Yat-sen University
c.ziliang@yahoo.com, z.yongsensmile@gmail.com, {laizhr,Gql}@jnu.edu.cn, linliang@ieee.org

Abstract

Invariant representation learning (IRL) encourages the prediction from invariant causal features to labels deconfounded from the environments, advancing the technical roadmap of out-of-distribution (OOD) generalization. Despite spotlights around, recent theoretical result verified that some causal features recovered by IRLs merely pretend domain-invariantly in the training environments but fail in unseen domains. The fake invariance severely endangers OOD generalization since the trustful objective can not be diagnosed and existing causal remedies are invalid to rectify. In this paper, we review a IRL family (InvRat) under the Partially and Fully Informative Invariant Feature Structural Causal Models (PIIF SCM /FIIF SCM) respectively, to certify their weaknesses in representing fake invariant features, then, unify their causal diagrams to propose ReStructured SCM (RS-SCM). RS-SCM can ideally rebuild the spurious and the fake invariant features simultaneously. Given this, we further develop an approach based on conditional mutual information with respect to RS-SCM, then rigorously rectify the spurious and fake invariant effects. It can be easily implemented by a small feature selection subnet introduced in the IRL family, which is alternatively optimized to achieve our goal. Experiments verified the superiority of our approach to fight against the fake invariant issue across a variety of OOD generalization benchmarks.

Introduction

A fundamental pressumption of machine learning widely believes that models are trained and tested with samples identically and independently (*i.i.d.*) drawn from a distribution. Whereas in practice, the models are inevitably trained and deployed in ubiquitous scenarios so that they poorly perform than what was expected, due to the violation of the *i.i.d.* condition inducing *distributional shift* across various scenarios. The failure could be understood from the view of representation learning, where the *i.i.d.* condition typically achieves the feature generalizing in a distribution, but unfortunately, at the sacrifice of generalization beyond this observed distribution. It is obviously impossible to obtain the domain universe by collecting data from all scenarios, so how to ***learn reprsentation*** with limited observed domains for chasing the ***invariant performance to unseen domains***, have gradually become the promising trend known as invariant representation learning (IRL) for out-of-distribution (OOD) generalization (Shen et al. 2021; Wang et al. 2022a).

The emergence of IRL dates back to approaches for domain adaptation (Ganin et al. 2016; Zhao et al. 2019) where data drawn from a test domain (so-called target domain) can be accessed to quantify the distributional shift, thus, invariant representation is spontaneously obtained while minimizing the domain shift. In the OOD generalization setup, only a few number of domains are available whereas the goal turns to learning the invariant representation to unseen domains. It becomes more challenging since minimizing the observed domain gaps does not imply the model generalization to unseen domains. The recent development of causal inference (Peters, Bühlmann, and Meinshausen 2016; Mahajan, Tople, and Sharma 2021) provided a set of innovative principles, *i.e.*, Invariant Causal Prediction (ICP), of connecting the IRL and OOD generalization. Most IRL frameworks henceforth consider data as an endogenous vairable generated through a Structural Causal Model (SCM) (Pearl 2010), which could be partitioned into different *environment factors* where each one corresponds to a specific intervention action taken in the SCM. In such regards, IRL aims for the recovery of invariant features via the arbitrary environment interventions for diminishing spurious correlation with the label. Of particular prominent methods are Invariant Risk Minimization (IRM) (Arjovsky et al. 2019), Invariant Rationalization (InvRat) (Chang et al. 2020; Li et al. 2022), REx (Krueger et al. 2021) and some other approaches in the similar spirit (Zhou et al. 2022; Ahuja et al. 2020; Li et al. 2022). Their objectives are optimized to prevent the classifier from overfitting to environment-specific properties.

Fake OOD Invariant Effect

Despite the potential and popularity of IRL, plentiful follow-up studies unveiled IRLs' unreliability to learn invariant representation (Kamath et al. 2021; Nagarajan, Andreassen, and Neyshabur 2020; Rosenfeld, Ravikumar, and Risteski 2020), in which the most notorious problem is probably the ***fake invariant effect*** . Particularly, given each environment factor to identify a specific spurious feature in a SCM, if the number of latent environment factors less than the capacity of spurious features, latent spurious correlation would pretend as an invarant part of the algorithm-recovered features recov-

[*]indicate corresponding author

ered by IRL. The problem arouses from the existence of underlying shortcut $\Phi(\cdot)$ between invariant causal features $Z_{\rm c}$ and spurious features $Z_{\rm s}$. It receives the spurious variable to endow $[Z_{\rm c}, \Phi(Z_{\rm s})]$ with the invariant property across training environments, where the classifier prefers $[Z_{\rm c}, \Phi(Z_{\rm s})]$ rather than $Z_{\rm c}$ for IRL. While the OOD generalization easily fails since $Z_{\rm s}$ depends on environments that allows arbitrary change during testing.

The fake invariance typically rises from the scarcity of environments that implicitly raises the "degree of freedom" of invariant representation. The uncontrolled "degree of freedom" are observed both in the linear and non-linear cases, where several recent efforts attempted to recover the true invariant features through the lens of causality. However, existing paradigms fail to incorporate $\Phi(Z_{\rm s})$ as a part of SCM. It endangers OOD generalization since no knowledge of the data assumption on the underlying environments may cause a paradox for IRL (Ahuja et al. 2021).

Contributions

To solve the problem above, our work provides the first rigorous investigation of considering the fake variant shortcut $\Phi(Z_{\rm s})$ as a latent variable across diverse SCM data assumptions. Specifically, we firstly investigated two famous SCM data assumptions (Partially and Fully Informative Invariant Feature, PIIF SCM and FIIF SCM) commonly employed by existing IRL frameworks (Ahuja et al. 2021). Under the background of a IRL family derived from (Chang et al. 2020; Li et al. 2022), we certify PIIF SCMs impossibly to incorporate $\Phi(Z_{\rm s})$ whereas the paradigm of FIIF SCM surprisingly suits the information-theoretic properties behind $\Phi(Z_{\rm s})$. To obtain the best of both worlds, we propose a novel ReStructured SCM framework combing PIIF SCM and FIIF SCM to simultaneously rebuild and isolate the spurious and the fake invariant charateristics.

Given this, we further proved why the IRL family only recovers the label-dependent spurious features but fails to mitigate the fake invariant features under the RS-SCM framework, and propose a conditional mutual information objective to rectify the negative invariant effect caused by $\Phi(Z_{\rm s})$. It can be easily implemented by a subnetwork to select invariant features then merge with the IRL family, which are alternatively trained to prevent invariant representation from the fake invariant effects. Diagnostic experiments and five large-scale real-world benchmarks validates our work.

Related Work

OOD generalization or domain generalization investigate the principles to extend the empirical risk minimization (ERM) to suit the data beyond the training distributions (Wang et al. 2022a; Shen et al. 2021). Before IRL becoming the trend, there have been three famous research lines. *Data augmentation* increases the diversity of observed domains by taking complex operations to transform training data, *i.e.*, randomization, mixup, altering location, texture and replicating the size of objects (Khirodkar, Yoo, and Kitani 2019; Wang, Li, and Kot 2020; Yu et al. 2023), etc. *Meta-learning* optimizes a general domain-agnostic model, which turns into a domain-specific version with a few of the domain-specific samples for the test adaptation (Shu et al. 2021; Chen et al. 2023). Ensemble approaches integrated submodels with regards to diverse training domains to generalize unseen distributions (Lee, Kim, and Kwak 2022; Chu et al. 2022).

Massive OOD generalization literatures are deeply related with IRL. The rationale behind aims to minimize the upper bounds of the prediction errors in unseen distributions which used to rely upon the covariate shift presumptions, yet (Chen and Bühlmann 2021; Kuang et al. 2018) implausible when the spurious correlation occurs. Increasing attentions were repayed to causality to tackle the issue. Inspired by ICP (Peters, Bühlmann, and Meinshausen 2016), a plenty of IRLs treat the predictions as invariant factors across different domains, which recovers the causation from feature to label regardless of environment interventions (Arjovsky et al. 2019; Ahuja et al. 2020; Chang et al. 2020; Krueger et al. 2021; Li et al. 2022; Jiang and Veitch 2022). Some causal learning achieve OOD generalization beyond ICP (Jalaldoust and Bareinboim 2023; Wang et al. 2022b; Lv et al. 2022).

Recent critics are discussed to this roadmap due to the unsatisfied recovery of invariant features. IRMs were denouced since its feasible variant IRMv1 poorly adapts to deep models (Zhou et al. 2022) and lacks the robustness of environment diversity (Huh and Baidya 2022; Lin et al. 2022). (Nagarajan, Andreassen, and Neyshabur 2020) uncovered two failure modes caused by geometric and statistical skews in their nature. (Rosenfeld, Ravikumar, and Risteski 2020) rigorously exhibited the fake invariance in the linear setting and empirically reported the issue over a wide range of IRL methods. Despite Invariant Information Bottleneck (IIB) (Li et al. 2022) claiming their capability to solve this issue, our causal diagnosis inspired by (Ahuja et al. 2021) verified that their solution powerless of the fake invariance.

Our work is closely related with Pearl's causality, the advanced knowledge (Pearl 2010, 2009) before reading.

Causal Diagnosis for Fake Invariance

We first review IRL and its fake invariant issue, then propose a new SCM to reflect the fake invariance in causal diagram.

Preliminary

OOD Generalization & IRL setup. Suppose we are given datasets $\mathcal{D}$=$\{\mathcal{D}_e\}$ for training, where each one refers to the training environment $e \in \mathcal{E}_{\rm tr}$ collected from the environment universe $\mathcal{E}$ ($\mathcal{E}_{\rm tr}$⊂$\mathcal{E}$). For a training set $\mathcal{D}_e = \{\boldsymbol{x}_i^e, \boldsymbol{y}_i^e\}_{i=1}^{n_e}$, each sample with its label was *i.i.d* drawn from the underlying joint distribution $\mathcal{P}_e$. The purpose of OOD generalization is to learn a model f with $\mathcal{D} = \{\mathcal{D}_e\}_{e \in \mathcal{E}_{\rm tr}}$ for enabling the label prediction to the samples drawn from arbitrary environments in $\mathcal{E}$, thus, minimizing the population risk as follows:

$$\mathcal{R}_{\mathcal{E}}(f) = \max_{e\sim\mathcal{E}}[\mathcal{R}^e(f)] = \max_{e\sim\mathcal{E}}\big[\mathbb{E}_{\mathcal{P}_e(\boldsymbol{x},\boldsymbol{y})}\big[\mathcal{L}\big(f(\boldsymbol{x}), \boldsymbol{y}\big)\big]\big], \quad (1)$$

where $\mathcal{L}(\cdot,\cdot)$ denotes the loss function with regards to a task; $f(\boldsymbol{x})$=$\rho(h(\boldsymbol{x}))$ in which $h(\boldsymbol{x})$ denotes the feature extracted from $\boldsymbol{x}$ by encoder h and ρ receives $h(\boldsymbol{x})$ to predict the label.

Obviously the population risk in Eq.1 is impossible to directly approximate since we have distributions of $\mathcal{E}_{\rm tr}$ instead of $\mathcal{E}$ during training. IRLs resort to training the model f over

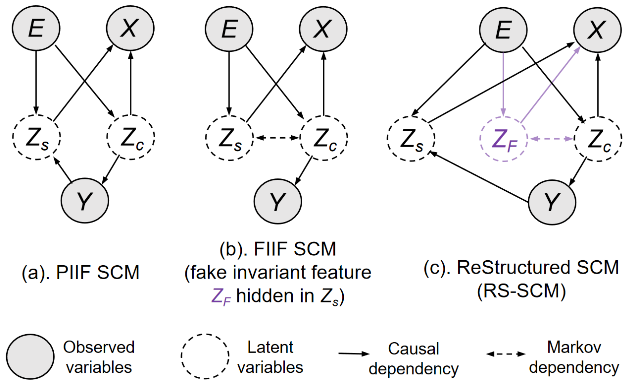

Figure 1: SCMs for InvRat and IIB in distribution shifts:(a). SCM with the PIIF condition ($Y \not\perp E$,$Z_s|Z_c$); (b). SCM with the FIIF condition ($Y \perp E$,$Z_s|Z_c$) where fake invariant features Z_F=$\Phi(Z_s)$ hide; (c).RS-SCM incorporates the FIIF and PIIF with the fake invariant variable Z_F.

$\mathcal{D}$ to find the representation with the domain-invariant properties. For instance, InvRat and its derivation IIB (Li et al. 2022) achieve this goal by maximizing the mutual information (MI) in the usage of the invariance contraint:

$$\max_{\rho,h} I[Y;h(X)] \textbf{ s.t. } Y \perp E \mid h(X), \tag{2}$$

where the capital letters denote variables and $\perp$ denotes their probabilistic independence. $I[Y;h(X)]$ denotes the MI between Y (label variable) and $h(X)$ (feature variable) extracted from X (data variable). The invariance contraint equals to minimze the Conditional MI (CMI) $I[Y;E|h(X)]$ between Y and E (environment variable) under the condition $h(X)$. They could be jointly formulated and optimized through their variational upper bounds (Alemi et al. 2016).

Existing work mostly investigates IRL from the IRM perspective, while our paper focuses on the IRL paradigms derived from InvRat or IIB. It helps to avoid many optimization issues that conventionally occur in IRM and its variants.

Causal Rationales behind IRLs. It is known that ***OOD generalization is possible under practical causal assumptions***, representing the data generation process with different types of distributional shifts with regards to different environment interventions (Ahuja et al. 2021). Resembling the similar principles, we provide a latent-variable SCM perspective to observe how InvRat / IIB generates X from the latent variable Z. It is concatenated by the invariant features Z_c and the variant counterpart Z_s with respect to the environment drawn from E. In the regards of latent interaction between Z_c and Z_s, we may further categorize the SCMs into the types *Fully Informative Invariant Features* (FIIF) and *Partially Informative Invariant Features* (PIIF), depending upon whether Z_c is fully informed by Y, *i.e.*, $Y \perp E, Z_s \mid Z_c$. Formal definitions are provided without additive noise for simplicity:

Assumption 1 (PIIF Structural Causal Model (SCM))**.**

$$\begin{aligned} Y &:= f_{\text{inv}}(Z_c), & Z_c &:= f_{\text{env}}(E), \\ Z_s &:= f_{\text{spu}}(E, Y), & X &:= f_{\text{gen}}(Z_c, Z_s); \end{aligned}$$

Assumption 2 (FIIF Structural Causal Model (SCM))**.**

$$\begin{aligned} Y &:= f_{\text{inv}}(Z_c), & Z_c &:= f_{\text{env}}(Z_s, E), \\ Z_s &:= f_{\text{spu}}(Z_c, E), & X &:= f_{\text{gen}}(Z_c, Z_s). \end{aligned}$$

Figure.1.(a-b) visualize the causal diagrams under the assumptions for InvRat and IIB. Despite representing different classes of distributional shifts, PIIF and FIIF SCMs simultaneously consider Z_s as the spurious correlation of the causal routine from data X to label Y, then, seeking to deconfound the spurious factors from E. Their common goal is to learn the feature encoder $h(\cdot)$ for recovering Z_c, then facilitate the invarint causal prediction $Z_c \overset{f_{\text{inv}}}{\rightarrow} Y$. It is highlighted that the previous work discussed for InvRat /IIB were almost derived from the PIIF Assumption (Figure.1.(a)). However, the FIIF Assumption (Figure.1.(b)) was seldom investigated because spurious correlations mostly live in the situations where Z_s is partially informed by Y ($Y \not\perp E$,$Z_s|Z_c$), *e.g.*, the spurious features refer to visual background information in an image. In contrast, our work prefers to the necessity of incorporating the FIIF SCM Assumption since *the fake invariant effects are probably hidden in the FIIF-SCM Assumption.*

Fake Invariance from A Causal Lens

To verfiy our claim, we need to overview the concept of fake invarian features. As demonstrated by (Li et al. 2022), they are some spurious features Z_s that pretend to be the domain-invariant part of Z by the shortcut $\Phi(\cdot)$ from two viewpoints

- $Y \perp E|\ Z_c$,$\Phi(Z_s)$ (**Fake invariance**): Combining Z_c and $\Phi(Z_s)$ can produce the lower empirical risk than the invariant risk, implying the domain-invariant property.
- $Y \not\perp E|\ \Phi(Z_s)$ (**Spuriousness**): For arbitrary Z_s,Y,E under the SCM Assumptions aforementioned, we discover $Y \not\perp E|\ Z_s \leftrightarrow I[Y;E|Z_s]>0$. Consider the shortcut $\Phi()$ effect on $I[Y;E|Z_s]$ reduced to

$$\begin{aligned} & I[Y;E|\Phi(Z_s)] \\ =& I[Y;E] - \big(I[Y;\Phi(Z_s)] - I[Y;\Phi(Z_s)|E]\big) \\ \geq& I[Y;E] - \big(I[Y;Z_s] - I[Y;Z_s|E]\big) \\ =& I[Y;E|Z_s] > 0, \end{aligned}$$

  which leads to such property.

Given these, we reconsider the environment interventions in the PIIF and FIIF SCM Assumptions, respectively, then discuss whether $\Phi(Z_s)$ embedded in their frameworks. Notice that $\Phi(\cdot)$ is solely the pathway algorithmic-recovered from the feature encoder $h(\cdot)$. It does not refer to any dependency under SCM assumptions.

PIIF SCM Fail to Identify the Fake Invariance. In the PIIF SCM Assumption, some evidences in Figure.1(b) can be readouted through the d-seperation rules (Pearl 2010):

- $Y \not\perp E$ implies the label marginal alters according to the environment interventions (the *non i.i.d.* property);
- $Y \perp E\ |Z_c$ demonstrates the independence between the label and the environment intervention provided with Z_c. So $h(\cdot)$ recovers the invariant features from Z_c and may achieve OOD generalization regardless of E;

- $Y \not\perp E\,|Z_s$ and $Y \not\perp E|Z_c,Z_s$ respectively demonstrate that (1). Z_s indicates the non-invariant property that we mentioned; (2). If $h(\cdot)$ recovers Z_c, Z_s simultaneously, the independence between Z_c and Y will not hold since Z_s works for a collider between E and Y.

The observations explain a broad range of questions of IRLs except for the fake invariant phenomenon. Specifically, such issue is typically caused by the spurious features Z_s via the shortcut $\Phi(Z_s)$, whereas $\Phi(Z_s)$ can not be reflected by Z_s in the PIIF SCM since evidences $Y \not\perp E\,|Z_s$ and $Y \not\perp E|Z_c,Z_s$ prevent the encoder $h(\cdot)$ from recovering features in Z_s. But what if $\Phi(Z_s)$ indicates a part of Z_c? This conjecture is also impossible in the PIIF SCM setup due to the *spuriousness* of $\Phi(Z_s)$: $Y \not\perp E|\,\Phi(Z_s)$ obviously conflicted with the domain-invariant property required for the features in Z_c.

FIIF SCM Implies the Fake Invariance. Provided with the failure witnessed in the PIIF SCM Assumption, we turn to the FIIF SCM Assumption and verify why the fake invariant features could be distinctly denoted as Z_s in Figure.1(b). We analyze the causal independences behind the FIIF SCM Assumption by following the same routine of the PIIF SCM Assumption. Such data distribution satisfies

- $Y \not\perp E$ and $Y \perp E\,|Z_c$ both hold as the FIIF SCM does;
- $Y \not\perp E\,|Z_s$ indicates Z_s's the spurious nature;
- $Y \perp E\,|Z_c,Z_s$ demonstrates that combing Z_c and Z_s leads to the invariant representation, however, Z_s should not be included since it implies spurisous correlations.

Observe that the second property suggests Z_s being domain-specific for the label prediction, whereas combining Z_c and Z_s results in the domain-invariant property that intervents the causal prediction over $Z_c,Z_s \overset{f_{\text{inv}}}{\to} Y$, which should have been $Z_c \overset{f_{\text{inv}}}{\to} Y$ instead. In this case, the second and the third observations for Z_s in the FIIF SCM setup do exactly refer to the *fake invariance* and the *spuriousness* behind $\Phi(Z_s)$.

ReStructued SCM. Despite incorporating the fake invariance, Z_s in the FIIF SCM Assumption contradicts the PIIF-SCM spurious correlation commonly found in practice. To obtain the best of both worlds, we restructure their SCMs and propose a new data generation regime that unify the PIIF and FIIF spurious correlations:

Assumption 3 (ReStructured SCM (RS-SCM, Figure.1(c)))**.**

$$Y := f_{\text{inv}}(Z_c), \quad Z_c := f_{\text{env}}(E, Z_F), \quad Z_s := f_{\text{spu}}(E, Y),$$
$$Z_F := f_{\text{fake}}(E, Z_c), \quad X := f_{\text{gen}}(Z_c, Z_s);$$

The RS-SCM extends the previous PIIF SCM by branching the Z_s-based spurious features to embrace the Z_F as our fake invariant features *i.e.*, Z_F=$\Phi(Z_s)$. The independencies between Z_F and the other variables keep consistent with the spurious features Z_s used in the FIIF SCM Assumption. Notably, the fake invariant effect only happens while the training environments overloaded with all spurious factors, therefore the causal subgraph with respect to Z_F in the RS-SCM should be adaptively deactivated beyond this situation. The switchable SCM mechanism is inspired from the heterogeneous causal graph (Watson et al. 2023), where we highlight the switchable parts by purple in Figure.1(c).

Remark 1. *The RS-SCM Assumption concurrently embeds spurious features and fake invariant features.*

Methodology

In this section, we elaborate our methodology derived from the restructured causality. We first review the strategies in InvRat and IIB, then, showing how they fail to rectify $\Phi(Z_s)$. Then we formulate our rectification objective to calibrate InvRat and IIB in an invariant learning manner.

InvRat Family Does Not Rectify $\Phi(Z_s)$

In the InvRat family, the vanilla InvRat obviously fails due to no effort paid to rectify $\Phi(Z_s)$ by Eq.2. Its derivation IIB advocates the minmal information between X and $h(X)$, *i.e.*, $\min_h I[X; h(X)]$. The constraint pernalizes the capacity of invariant feature recovery in $h(X)$, then combined with the invariant constraint $\min_h I[Y; E|h(X)]$. It was deemed to remove $\Phi(Z_s)$ hidden in the recovered feature $h(X)$, which is unreliable since their analysis is built upon the PIIP SCM Assumption where $\Phi(Z_s)$ can not be reflected by their latent variables. But under our RS-SCM Assumption, whether the constraint $\min_h I[X; h(X)]$ enables the fake invariance elimination? Our theoretic result also denies such guess:

Proposition 1. *In the RS-SCM Assumption, given invariant feature $Z_c \in \mathbb{R}^{n_c}$ and fake invariant features $\Phi(Z_s) \in \mathbb{R}^{n_F}$ as a feature subset of fake invariant variable Z_F: $\Phi(Z_s) \subset Z_F$, we can find $Z \in \mathbb{R}^{n_c}$ with $Z \cap \Phi(Z_s) \neq \varnothing$ that satisfies*

$$\begin{aligned} &\lambda I[Y; E|Z] + \beta I[X; Z] \\ &\leq \lambda I[Y; E|Z_c \cup \Phi(Z_s)] + \beta I[X; Z_c \cup \Phi(Z_s)], \\ &\qquad\qquad\qquad\qquad \textbf{\textit{s.t.}} \ \ \forall \lambda, \beta \in \mathbb{R}^+. \end{aligned} \tag{3}$$

The justification elaborates that for each invariant feature Z_c recovered by $h()$, the FIIP SCM may search the feature Z in the identical latent space of Z_c to bound the invariant contraint of the IIB strategy, however, Z satisfies $Z \cap \Phi(Z_s) \neq \varnothing$ wherein the fake invariant features might be included by this reprsentation. IRLs conventionally optimize their models by way of non-convex variational bounds, thus hardly to certify whether training $h(\cdot)$ may result in the recovery of Z or Z_c. In terms of Proposition.1 and what we previously discussed, the conlusion is drawn to the InvRat family in RS-SCM:

Remark 2. *Under the RS-SCM Assumption, InvRat and IIB strategies distinguish the spurious features Z_s whereas fail to elimiate the fake invariant features $Z_F = \Phi(Z_s)$.*

Remark.2 illustrates the bright side of the InvRat familiy: the ability to debias $Z_c,Z_F \overset{f_{\text{inv}}}{\to} Y$ from the spurious factors Z_s by the invariant independence contraint $\big(i.e., Y \perp E\,|h(X)\big)$. To achieve OOD generalization, we are required to prevent existing InvRat variants from the unexpected recovery of Z_F.

Rectification Approach by RS-SCM

We move foward our discussion of how to wipe out Z_F from $Z_c \cup Z_F$. Under the RS-SCM Assumption, we reconsider the Markov dependency across Z_c and Z_F then distinguish them according to their different behaviors for the label prediction conditioned with each other and Z_s. Specifically, when provided with Z_F and Z_s, the d-seperation principle judges the causal prediction $Z_c \overset{f_{\text{inv}}}{\to} Y$ with $Y \not\perp Z_c | Z_F, Z_s$. It implies the causal path activated to maximize the CMI:

$$\max_{Z_c} I[Y; Z_c|Z_F, Z_s] = \max_{Z_c} I[Y; Z_c|h(X)/Z_c], \quad (4)$$

which is optimized for recovering Z_c from invariant encoder $h(X)$ learned by InvRat or IIB.

Similarly, we observe the causal dependency across Y and Z_F, then analyze the label prediction $Z_F \overset{f_{\text{env}}}{\to} Z_c \overset{f_{\text{inv}}}{\to} Y$ given Z_s,Z_c as the condition. It refers to $Y \perp Z_F | Z_c, Z_s$ that equivalently minimizes the CMI constraint:

$$\min_{Z_F} I[Y; Z_F|Z_c, Z_s] = \min_{Z_F} I[Y; Z_F|h(X)/Z_F]. \quad (5)$$

The contraint above helps us to identify Z_F from $h(X)$.

Note that when $h(X)$ has been well trained by the InvRat or the IIB, their models encourage $h^*(X)$=$Z_F \cup Z_c$. The nice property unifies Eq.4 and Eq.5 into the same objective, *i.e.*,

$$\begin{aligned} &\min_{Z_c, Z_F} I[Y; Z_F|h^*(X)/Z_F] - \lambda I[Y; Z_c|h^*(X)/Z_c] \\ &= \min_{Z_c} I[Y; h^*(X)/Z_c|Z_c] - \lambda I[Y; Z_c|h^*(X)/Z_c]. \end{aligned} \quad (6)$$

where λ indicates the trade-off co-efficient. Maximizing and minimizing CMI are intractable while thanks to the symmetry between Eq.4 and Eq.5, the CMI decomposition holds as

$$\begin{aligned} I[Y; h^*(X)/Z_c|Z_c] &= -H(Y|h^*(X)) + H(Y|Z_c); \\ I[Y; Z_c|h^*(X)/Z_c] &= -H(Y|h^*(X)) + H(Y|h^*(X)/Z_c), \end{aligned}$$

so that we simplify Eq.6 for the Z_c recovery from $h^*(X)$:

$$\begin{aligned} &\min_{Z_c} \ I[Y; h^*(X)/Z_c|Z_c] - \lambda I[Y; Z_c|h^*(X)/Z_c] \\ &= \min_{Z_c} \ H(Y|Z_c) - \lambda H(Y|h^*(X)/Z_c) + (\lambda - 1)H(Y|h^*(X)) \end{aligned} \quad (7)$$

where $(\lambda - 1)H(Y|h^*(X))$ is constant in the optimization. The objective implies that rectification only needs to select features from $h^*(X)$ to improve the causal invariant prediction $Z_c \overset{f_{\text{inv}}}{\to} Y$ (the first term) and discourage the fake invariant prediction $Z_F \overset{f_{\text{env}}}{\to} Z_c \overset{f_{\text{inv}}}{\to} Y$(the second term). The joint CMI nature behind Eq.7 holds the theoretical ganrantee as

Proposition 2. *Suppose that $h^*(X)$=$Z_F \cup Z_c$ under the RS-SCM Assumption. If feature Z recovered from $h^*(X)$ satisfies $I[Y; h^*(X)/Z|Z]$=0 and $I[Y; Z|h^*(X)/Z]$>0, it holds $Z = Z_c$ or $Z = Z_c \cup Z_F$.*

The proposition implies the rectification may lead to the ideal Z=Z_c or the trivial result that collapses into $Z_c \cup Z_F$. To prevent the trivial solution, we encourage the joint CMI objective optimized along with $\max_Z I(Z; h*(X)/Z)$, where $I(Z; h*(X)/Z)$>0 helps to get rid of the collapse.

Interplay Invariant Learning

Given the rectification approach by Eq.7, we propose a novel framework for OOD generalization.

Neural Soft-Feature Selector. Eq.7 demands a small network that selects features from $h(X)$ to recover Z_c. We take a simple two-layer architecture then adjust the scale complexity according to the tasks involved. The sub-network $s(\cdot)$ receives the latent layer's output from $h(X)$ to make the soft feature selection on $h(X)$. Specifically, $s(h(X))$ goes through a series of sigmoid activation functions to yield a vector with the same dimension of $h(X)$, where each positive output means that the feature is selected as Z_c. So $s(h(X)) \odot h(X)$ corresponds to the soft-feature selection for Z_c and $\big(\mathbf{1} - s(h(X))\big) \odot h(X)$ corresponds to the soft-feature selection for Z_F ($\odot$ indicates the entry-wise product). The objective Eq.7 turns into

$$\begin{aligned} \min_{s} \ &\mathbb{E}\big[\mathcal{L}\big(Y, s(h(X)) \odot h(X)\big)\big] \\ &- \lambda \mathbb{E}\big[\mathcal{L}\big(Y, (\mathbf{1} - s(h(X))) \odot h(X)\big)\big] \end{aligned} \quad (8)$$

where we take the task-specific loss to approximate the conditional entropy,*i.e.*, $\mathbb{E}[\mathcal{L}(Y,Z)] \to H(Y|Z)$.

Framework. We show how to combine Eq.8 with InvRat to learn invariant representation alternatively. Derived from the variational upper bounds of Eq.2, the InvRat family plays an adversarial game to jointly train three sub-networks, *i.e.*, feature encoder $h(\cdot)$, invariant predictor $g_{\text{i}}(\cdot)$, domain-aware predictor $g_{\text{d}}(\cdot)$:

$$\begin{aligned} \min_{h, g_{\text{i}}} \max_{g_{\text{d}}} \ &\mathbb{E}\big[\mathcal{L}(Y; g_{\text{i}}(h(X)))\big] \\ &+ \beta\big(\mathbb{E}\big[\mathcal{L}(Y; g_{\text{i}}(h(X)))\big] - \mathbb{E}\big[\mathcal{L}(Y; g_{\text{d}}(h(X)))\big]\big) \end{aligned} \quad (9)$$

Given subnetworks pre-trained by Eq.9, our invariant learning framework alternatively performs to (1). train the feature selector $s(\cdot)$ with respect to Eq.8 in the outer loop; (2). finetune the InvRat subnetworks by Eq.9 in the inner loop. It is illustrated in Figure.2.

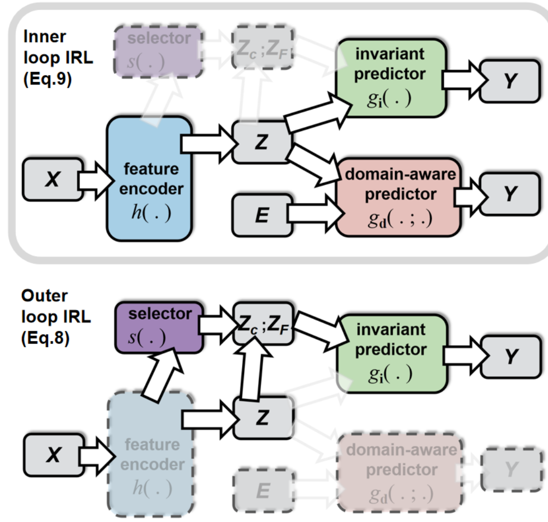

Figure 2: Our interplay invariant learning (IIL) framework. The transparency implies the network frozen or the variables not activated in this alternative phase (Best viewed in color).

Experiments

In this section, we firstly conduct the diagnostic experiments on the benchmarks derived from recent studies (Arjovsky et al. 2019; Ahmed et al. 2020) broadly applied in IRL for OOD generalization. It aims to validate whether our IIL can (1). rectify the fake invariant effect as demonstraed by our theoretical analysis; (2). remain the capability of InvRat and IIB to learn invariant representation. Afterwards, we evaluate our IIL framework in five competitive benchmarks for domain generalization in the wild, in order to verify IIL's feasibility in complex scenarios. Notice that InvRat is originally proposed for rationalization but IIB exactly share the most of its optimization pipelines beyond the MI constraint $I[X; h(X)]$. In this regards, our experiments consider InvRat as the IIB without this regularization.

Benchmarks. The diagnostic study provides the forensic of IRL baselines under the RS-SCM Assumption. It requires the datasets generated by the same causal mechanism, however, existing diagnostic benchmarks are generated by either FIIF or PIIF SCM, hardly fullfiling our demand (Arjovsky et al. 2019; Ahmed et al. 2020). We observe that RS-SCM consists of FIIF and PIIF SCM Assumptions so that combine their generation recipes to build our diagnostic benchmark to evaluate the invariant learning quality.

Dataset	Z_c	Z_s	Z_F	Training/Test Samples
CS-MNIST-CIFAR	Digit	CIFAR	Color	
CS-MNIST-COCO	Digit	Object	Color	

Table 1: The summary of our diagnostic benchmarks.

Specifically, we consider the ten-class digit classification mission derived from the CS-MNIST (FIIF) benchmark in (Ahuja et al. 2021). It consists of two environments for training with 20,000 samples each and one environment for evaluation with the same number of data. In this setup, Z_c would refer to the shape of digit and ten digit classes would be associated with ten colors respectively, with an environment-specific probability p_e. The color indicates the fake invariant variable Z_F and the association indicates the Markov dependency between Z_c and Z_F, and p_e implies the environment dependences from E to latent variables Z_F and Z_c: p_e indicates their association activated otherwise the digit would randomly associated with the ten colors.

The generated digits can not represent the RS-SCM since the spurious factor Z_s has not been included. In this case, we resemble the composition rule in CIFAR-MNIST (Zhou et al. 2022) whereas we classify MNIST instead of CIFAR. So given each colored digit generated by the previous strategy, we combine it with a CIFAR image drawn from the generative process following the PIIF SCM Assumption (Arjovsky et al. 2019). Specifically, given a digit generated by the previous process, we take a random flip with 25% chance to randomly change its label; then we associate this digit-class label with a CIFAR class with environment-dependent probability $\hat{p}_e$. So we have the CS-MNIST-CIFAR to represent the RS-SCM Assumption where the spurious vairable Z_s is indicated by the CIFAR classes. We replay this process with Color-COCO then get the second RS-SCM benchmark

Methods	ID Acc (↑)	OOD Generalization Acc (↑)		
	No shift	Z/Z_s	Z/Z_F	$Z/(Z_s, Z_F)$
ERM	93.22	11.87	13.76	10.13
IRM	94.49	59.58	53.43	50.06
IRM+IB	96.14	66.98	70.42	59.71
InRav	89.25	63.39	65.75	60.89
IIB	**96.76**	**70.98**	69.42	66.23
InvRav(+ours)	91.72 ↑	64.47↑	69.23 ↑	64.40 ↑
IIB(+ours)	95.17↓	70.19↓	**70.54** ↑	**68.37** ↑

Table 2: ID / OOD generalization accuracies on CS-MNIST-CIFAR. Z/Z_s, Z/Z_c, and $Z/(Z_s, Z_c)$ indicate different distributional shifts between training and test (without spurious factor, without fake invariant, without the both).

Methods	ID Acc (↑)	OOD Generalization Acc (↑)		
	No shift	Z/Z_s	Z/Z_F	$Z/(Z_s, Z_F)$
ERM	92.63	10.24	11.47	9.67
IRM	94.49	49.67	54.26	47.19
IRM+IB	**96.14**	56.91	63.92	55.27
InRav	89.25	53.07	61.75	51.33
IIB	92.44	61.14	66.38	57.62
InvRav(+ours)	91.72 ↑	58.86 ↑	66.42 ↑	56.16 ↑
IIB(+ours)	93.17 ↑	**63.72** ↑	**69.59** ↑	**62.35** ↑

Table 3: ID / OOD generalization accuracies on CS-MNIST-COCO. Z/Z_s, Z/Z_c, and $Z/(Z_s, Z_c)$ indicate different distributional shifts between training and test (without spurious factor, without fake invariant, without the both).

CS-MNIST-COCO (see Table.1). $p_e = 1, 0.9$ and $\hat{p}_e = 1, 0.9$ are set up for two training environments, respectively.

Beyond the diagnostic datasets, we also conduct the experiments on VLCS, PACS, Office-HOME, Terra-Incognita and DomainNet, which refer to DomainBed (Gulrajani and Lopez-Paz 2020) for real-world OOD generalization.

Experimental setup. In terms of the feature encoder, invariant predictor and domain-aware predictor, we employs the architecutres applied in (Li et al. 2022). We take a simple two-layer network for CS-MNIST and a transformer-like subnetwork for DomainBed as our neural feature selectors.

Diagnostic OOD Generalzation

Baselines. Beyond InvRat and IIB, we take IRM (Arjovsky et al. 2019) and IRM+IB (Ahuja et al. 2021) implemented by IRMv1 variants as our IRL baselines. We also employed ERM as the borderline to judge the IRL performance.

Results. 20% training data are split into the validation set for CS-MNIST-CIFAR and CS-MNIST-COCO, where all baselines are evaluated to produce their in-distribution (ID) performances. Our evaluation is interested in OOD generalization across diverse distributional shifts: (1).Z/Z_s indicates the test environment with the spurious covariate shift (each test digit was randomly matched with an image drawn from CIFAR or ColorCOCO regardless of the image label); Z/Z_F indicates the test environment with the fake invariant distri-

	VLCS	PACS	Office-H	Incognita	DomainN	Average
ERM	77.2	83.0	65.7	41.4	40.6	61.6
IRM	**78.5**	83.5	64.3	**47.6**	33.9	61.6
VREx	78.3	84.9	66.4	46.4	33.6	61.9
CausalIRL	77.6	84.0	65.7	46.3	40.3	62.8
InRav	77.3	83.5	66.2	44.6	35.1	61.3
IIB	77.2	83.9	68.6	45.8	41.5	63.4
InvRav(+ours)	77.3	84.6	66.3	46.1	40.1	62.9
IIB(+ours)	77.6	**85.8**	**68.8**	**47.6**	**42.5**	**64.4**

Table 4: OOD generalization accuracy on DomainBed.

butional shift (each test digit was randomly matched with a color); $Z/(Z_s, Z_F)$ denotes the test environment containing the spurious and fake invariant distribution shifts concurrently (a test digit takes the both actions simultaneously). The accuracies across all baselines are evidenced in Table.2 (CS-MNIST-CIFAR) and Table.3 (CS-MNIST-COCO).

Our diagnostic experiment was conducted to address two major concerns to our approach. 1.(*identification concern*): if the fake invariance happens (*i.e.*, Z_F has been activated in the RS-SCM), why IRL needs to identify the fake invariant variable Z_z? 2.(*rectification concern*): whether our approach rectify the negative effect caused by the fake invariant features Z_z instead of other spurious covariates?

In view of the diagnosis concern, we investigate the comparison among diverse testbed scenarios. We first note that the ERM almost perform to approximate the random guess in arbitrary OOD situations, implying our diagnostic benchmark with diverse and significant distributional shifts. Beyond this, Z/Z_s arouses more severe accuracy drop compared with Z/Z_F. It makes sense since the introduced image contains spurious covariates with the higher dimensionality than the colorized pixels. But even so, the accuracies of $Z/(Z_s, Z_F)$ in the majority of baselines almost underperform their Z/Z_s counterparts. It verified that the covariate shift caused by the fake invariant variable Z_F could not be conveniently eliminated by addressing the shift caused by Z_s. It justifies the superiority of our approach, which identifies the covariate shift caused by Z_F to prevent the invariant prediction from the biased representaion (Z_s, Z_F).

Given this, we compare our approach with other baselines particularly, InvRat and IIB, to verify its rectification ability to the fake invariant factors Z_F. In Table.2, the performances of InvRat and IIB boosted by our IIL are inconspicuous in the ID scenarios and OOD scenario Z/Z_s (+2.48 in ID and +1.09 in Z/Z_s for InvRat, respectively; and receives negative transfer in ID and Z/Z_s for IIB), yet its accuracy boost significantly when comes to Z/Z_F and $Z/(Z_s, Z_F)$ scenarios *e.g.*, +3.48 in Z/Z_F and +3.51 in $Z/(Z_s, Z_F)$ for InvRat. In Table.3, the performance boost has been observed more sigficantly. The ablation evidences demonstrate that our rectification approach mainly works for eliminating the negative covariate shift caused by Z_F while thanks to the interplay learning manner, the overall performance get benefited.

Real-world OOD Generalization

Baselines. We follow the evaluation setup and testify all IRL baselines including ERM, IRM, VREx (Krueger et al. 2021), and the recent approach CausalIRL (Chevalley et al. 2022), which all belong to competitive IRL baselines. We also evaluated other 15 baselines apart from IRL approaches in Appendix.C. We employ the model selection strategy by leave-one-domain-out cross validation.

Results. In Table.4, we evaluate our approach by combining it with InvRat and IIB. In general, our approach improved InvRav by +1.6% and IIB by +1.0%. In terms of our theoretical finding and evidences shown in the diagnostic evaluation, we figure the improvement probably due to the fake variant factors eliminated by our approach. It verified that our approach is compatible with the information bottleneck regularization. Whereas we also observe that the increase in VLCS is very limited (+0.0 for InvRat and +0.4 for IIB in VLCS). It is probably due to VLCS composed by 5 classes across 4 datasets, thus, insufficient to capture the domain-specific complexity. Beyond this, our approach significantly outperform the other baselines.

Ablation

The ablation study has been provided in our ArXiv version.

Conclusion

This paper attempts to resolve the fake invariance problem for IRL, which undermines the OOD generalization performance. We proposes a novel structural causal model, Re-Structured SCM (RS-SCM) to reconstruct both spurious and fake invariant features from the data. It inspires a new approach to eliminate the spurious and fake invariant effects.

Acknowledgments

This work was supported in part by National Key R&D Program of China under Grant No.2021ZD0111601 and No.2022YFC3303603; in part by National Natural Science Foundation of China (NSFC) under Grant No.61836012, U21A20470, 62206110, 62077028, 62176103, and 62377028; in part by the Science and Technology Planning Project of Guangdong (2021B0101420003, 2020B0909030005, 2020B1212030003, 2020ZDZX3013, 2023ZZ03); in part by the Science and Technology Planning Project of Guangzhou (No.202206030007); in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003), in part by Jinan University and the Opening Project of Key Laboratory of Safety of Intelligent Robots for State Market Regulation (No.GQI-KFKT202205); in part by Guangdong Key Laboratory of Data Security and Privacy Preserving (Grant No. 2023B1212060036); in part by GuangDong Basic and Applied Basic Research Foundation under Grant No.2023A1515012845 and 2023A1515011374. Liang Lin is also leading the GuangDong Province Key Laboratory of Information Security Technology.

References

Ahmed, F.; Bengio, Y.; Van Seijen, H.; and Courville, A. 2020. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*.

Ahuja, K.; Caballero, E.; Zhang, D.; Gagnon-Audet, J.-C.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450.

Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*, 145–155. PMLR.

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. 2020. Invariant rationalization. In *International Conference on Machine Learning*, 1448–1458. PMLR.

Chen, J.; Gao, Z.; Wu, X.; and Luo, J. 2023. Meta-causal Learning for Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7683–7692.

Chen, Y.; and Bühlmann, P. 2021. Domain adaptation under structural causal models. *The Journal of Machine Learning Research*, 22(1): 11856–11935.

Chevalley, M.; Bunne, C.; Krause, A.; and Bauer, S. 2022. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*.

Chu, X.; Jin, Y.; Zhu, W.; Wang, Y.; Wang, X.; Zhang, S.; and Mei, H. 2022. DNA: Domain generalization with diversified neural averaging. In *International Conference on Machine Learning*, 4010–4034. PMLR.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Gulrajani, I.; and Lopez-Paz, D. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Huh, D.; and Baidya, A. 2022. The Missing Invariance Principle found–the Reciprocal Twin of Invariant Risk Minimization. *Advances in Neural Information Processing Systems*, 35: 23023–23035.

Jalaldoust, K.; and Bareinboim, E. 2023. Transportable Representations for Out-of-distribution Generalization. In *The ICML Workshop on Spurious Correlations, Invariance and Stability*.

Jiang, Y.; and Veitch, V. 2022. Invariant and transportable representations for anti-causal domain shifts. *Advances in Neural Information Processing Systems*, 35: 20782–20794.

Kamath, P.; Tangella, A.; Sutherland, D.; and Srebro, N. 2021. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, 4069–4077. PMLR.

Khirodkar, R.; Yoo, D.; and Kitani, K. 2019. Domain randomization for scene-specific car detection and pose estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1932–1940. IEEE.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 5815–5826. PMLR.

Kuang, K.; Cui, P.; Athey, S.; Xiong, R.; and Li, B. 2018. Stable prediction across unknown environments. In *proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1617–1626.

Lee, K.; Kim, S.; and Kwak, S. 2022. Cross-domain ensemble distillation for domain generalization. In *European Conference on Computer Vision*, 1–20. Springer.

Li, B.; Shen, Y.; Wang, Y.; Zhu, W.; Li, D.; Keutzer, K.; and Zhao, H. 2022. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7399–7407.

Lin, Y.; Zhu, S.; Tan, L.; and Cui, P. 2022. ZIN: When and How to Learn Invariance Without Environment Partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542.

Lv, F.; Liang, J.; Li, S.; Zang, B.; Liu, C. H.; Wang, Z.; and Liu, D. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8046–8056.

Mahajan, D.; Tople, S.; and Sharma, A. 2021. Domain generalization using causal matching. In *International Conference on Machine Learning*, 7313–7324. PMLR.

Nagarajan, V.; Andreassen, A.; and Neyshabur, B. 2020. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*.

Pearl, J. 2009. Causal inference in statistics: An overview.

Pearl, J. 2010. Causal inference. *Causality: objectives and assessment*, 39–58.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5): 947–1012.

Rosenfeld, E.; Ravikumar, P. K.; and Risteski, A. 2020. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*.

Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Shu, Y.; Cao, Z.; Wang, C.; Wang, J.; and Long, M. 2021. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9624–9633.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. 2022a. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, R.; Yi, M.; Chen, Z.; and Zhu, S. 2022b. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 375–385.

Wang, Y.; Li, H.; and Kot, A. C. 2020. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3622–3626. IEEE.

Watson, R. A.; Cai, H.; An, X.; McLean, S.; and Song, R. 2023. On Heterogeneous Treatment Effects in Heterogeneous Causal Graphs. *arXiv preprint arXiv:2301.12383*.

Yu, R.; Liu, S.; Yang, X.; and Wang, X. 2023. Distribution Shift Inversion for Out-of-Distribution Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3592–3602.

Zhao, H.; Des Combes, R. T.; Zhang, K.; and Gordon, G. 2019. On learning invariant representations for domain adaptation. In *International conference on machine learning*, 7523–7532. PMLR.

Zhou, X.; Lin, Y.; Zhang, W.; and Zhang, T. 2022. Sparse invariant risk minimization. In *International Conference on Machine Learning*, 27222–27244. PMLR.