LATENT CAUSAL INVARIANT MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Current supervised learning can learn spurious correlation during the data-fitting process, imposing issues regarding interpretability, out-of-distribution (OOD) generalization, and robustness. To avoid spurious correlation, we propose a **La**tent **C**ausal **I**nvariance **M**odel (LaCIM) which pursues *causal prediction*. Specifically, we introduce latent variables that are separated into (a) output-causative factors and (b) others that are spuriously correlated to the output via confounders, to model the underlying causal factors. We further assume the generating mechanisms from latent space to observed data to be *causally invariant*. We give the identifiable claim of such invariance, particularly the disentanglement of output-causative factors from others, as a theoretical guarantee for precise inference and avoiding spurious correlation. We propose a Variational-Bayesian-based method for estimation and to optimize over the latent space for prediction. The utility of our approach is verified by improved interpretability, prediction power on various OOD scenarios (including healthcare) and robustness on security.

1 INTRODUCTION

Current data-driven deep learning models, revolutionary in various tasks though, heavily rely on *i.i.d* data to exploit all types of correlations to fit data well. Among such correlations, there can be spurious ones corresponding to biases (*e.g.*, selection or confounding bias due to coincidence of the presence of the third factor) inherited from the data provided. Such data-dependent spurious correlations can erode the *(i)* interpretability of decision-making, *(ii)* ability of out-of-distribution (OOD) generalization, *i.e.*, extrapolation from observed to new environments, which is crucial especially in safety-critical tasks such as healthcare, and *(iii)* robustness to small perturbation (Goodfellow et al., 2014).

Recently, there is a Renaissance of causality in machine learning, expected to pursue causal prediction (Schölkopf, 2019). The so-called "causality" is pioneered by Judea Pearl (Pearl, 2009), as a mathematical formulation of this metaphysical concept grasped in the human mind. The incorporation of a *priori* about cause and effect endows the model with the ability to identify the causal structure which entails not only the data but also the underlying process of how they are generated. For causal prediction, the old-school causal learning frameworks (Peters et al., 2016; Bühlmann, 2018) causally related the output label Y to the *observed input* X, which however is NOT conceptually reasonable in scenarios with sensory-level observed data (*e.g. pixels in image classification*).

For such applications, we rather adopt the manner in Bengio et al. (2013); Biederman (1987) to relate the causal factors of Y to unobserved abstractions denoted by S, *i.e.*, $Y \leftarrow f_y(S, \varepsilon_y)$ via mechanism f_y. We further assume existence of additional latent components denoted as Z, that together with S generates the input X via mechanism f_x as $X \leftarrow f_x(S, Z, \varepsilon_x)$. Taking image classification as an example, the S and Z respectively refer to object-related abstractions (*e.g.*, contour, texture, color) and contextual information (*e.g.*, light, view). Such an assumption is similarly adopted in the literature of nonlinear Independent Components Analysis (ICA) (Hyvarinen and Morioka, 2016; Hyvärinen et al., 2019; Khemakhem, Kingma and Hyvärinen, 2020; Teshima et al., 2020) and latent generative models (Suter et al., 2019), which are *however* without separation of output-causative factors (*a.k.a*, S) and other correlating factors (*a.k.a*, Z) that can both be learned in data-fitting process.

We encapsulate these assumptions into a novel causal model, namely *La*tent *C*ausal *I*nvariance *M*odel (LaCIM) with causal graph illustrated in Fig. 1, in which we assume $f_x : S, Z \to X$, $f_y : S \to Y$ to be the *C*ausal *I*nvariant *Me*chanisms (CIMe) that hold under any circumstances with $\mathrm{P}(S, Z)$ allowed to be varied across domains. The incorporation of these *priories* can explain the spurious correlation

embedded in the back-door path from Z to Y (contextual information to the class label in image classification). To avoid learning spurious correlations, our goal is to identify the intrinsic CIMe f_x, f_y. Specifically, we first prove the identifiability (*i.e.*, the possibility to be precisely inferred) of the CIMe. Notably, far beyond the scope in existing literature (Khemakhem, Kingma and Hyvärinen, 2020), our results can implicitly, and are the *first* to disentangle the output-causative factors (*a.k.a*, S) from others (*a.k.a*, Z) for prediction, to ensure the isolation of undesired spurious correlation. Guaranteed by such, we propose to estimate the CIMe by extending the variational auto-encoder (VAE) (Kingma and Welling, 2014) to the supervised scenario. For OOD prediction, we propose to optimize over latent space under the identified CIMe. To verify the correctness of our identifiability claim, we conduct a simulation experiment. We further demonstrate the utility of our LaCIM via high explainable learned semantic features, improved prediction power on various OOD scenarios (including tasks with confounding and selection bias, healthcare), and robustness on security.

We summarize our contribution as follows: **(i) Methodologically**, we propose in section 4.1 a latent causal model in which only a subset of latent components are causally related to the output, to avoid spurious correlation and benefit OOD generalization; **(ii) Theoretically**, we prove the identifiability (in theorem 4.3) of CIMe f_x, f_y from latent variables to observed data, which disentangles output-causative factors from others; **(iii) Algorithmically**, guided by the identifiability, we in section 4.3 reformulate Variational Bayesian method to estimate CIMe during training and optimize over latent space during the test; **(iv) Experimentally**, LaCIM outperforms others in terms of prediction power on OOD tasks and interpretability in section 5.2, and robustness to tiny perturbation in section 5.3.

2 Related Work

The invariance/causal learning proposes to learn the assumed invariance for transferring. For the invariance learning methods in Krueger et al. (2020) and Schölkopf (2019), the "invariance" can refer to stable correlation rather than causation, which lacks the interpretability and impedes its generalization to a broader set of domains. For causal learning, Peters et al. (2016); Bühlmann (2018); Kuang et al. (2018); Heinze-Deml and Meinshausen (2017) assume causal factors as observed input, which is inappropriate for sensory-level observational data. In contrast, our LaCIM introduces latent components as causal factors of the input; *more importantly*, we explicitly separate them into the output-causative features and others, to avoid spurious correlation. Further, we provide the identifiability claim of causal invariant mechanisms. In independent and concurrent works, Teshima et al. (2020) and Ilse et al. (2020) also explore latent variables in causal relation. As comparisons, Teshima et al. (2020) did not differentiate S from Z; and Ilse et al. (2020) proposed to augment intervened data, which can be intractable in real cases.

Other works which are conceptually related to us, as a non-exhaustive review, include (i) transfer learning which also leverages invariance in the context of domain adaptation (Schölkopf et al., 2011; Zhang et al., 2013; Gong et al., 2016) or domain generalization (Li et al., 2018; Shankar et al., 2018); and (ii) causal inference (Pearl, 2009; Peters et al., 2017) which proposes a structural causal model to incorporate intervention via "do-calculus" for cause-effect reasoning and counterfactual learning; (iii) latent generative model which also assumes generation from latent space to observed data (Kingma and Welling, 2014; Suter et al., 2019) *but* aims at learning generator in the unsupervised scenario.

3 Preliminaries

Before proceeding, we introduce the problem setup and structural causal model Pearl (2009).

Problem Setup & Notation Let X, Y respectively denote the input and output variables. The training data $\{\mathcal{D}^e\}_{e\in\mathcal{E}_{\text{train}}}$ are collected from the set of multiple environments $\mathcal{E}_{\text{train}}$, where each e is with distribution $\mathrm{P}^e(X,Y)$ over $\mathcal{X}\times\mathcal{Y}$ and $\mathcal{D}^e = \{x_i^e, y_i^e, d^e\}_{i\in[n_e]} \overset{i.i.d}{\sim} \mathrm{P}^e$ with $[k] := \{1, ..., k\}$ for any $k \in \mathbb{Z}^+$. The $d^e \in \mathbb{R}^m$ denotes the one-hot encoded domain index for e, where $1 \le m := \left|\mathcal{E}_{\text{train}}\right| \le n := \sum_{e\in\mathcal{E}_{\text{train}}} n_e$. Our goal is to learn a model $f : \mathcal{X} \mapsto \mathcal{Y}$ that learns output-causative factors for prediction and performs well on the set of all environments $\mathcal{E} \supset \mathcal{E}_{\text{train}}$, which is aligned with existing OOD generalization works (Arjovsky et al., 2019; Krueger et al., 2020). We use respectively upper, lower case letter and Cursive letter to denote the random variable, the instance and the space, *e.g.*, a is an instance in the space $\mathcal{A}$ of random variable A. The $[f]_{\mathcal{A}}$ denotes the f restricted on dimensions of $\mathcal{A}$. The Sobolev space $W^{k,p}(\mathcal{A})$ contains all f such that $\int_{\mathcal{A}} \left|\partial_A f^\alpha\big|_{A=a}\right|^p da < \infty, \forall \alpha \le k$.

Structural Causal Model. The structural causal model (scm) is defined as the causal graph assigned with structural equations. The causal graph encodes the assumptions in missing arrows in a directed acylic graph (DAG): $G = (V, E)$ with V, E respectively denoting the node set and the edge set. The $Pa(k)$ denotes the set of parent nodes of V_k for each $V_k \in V$ and the $X \to Y \in E$ indicates the causal effect of X on Y. The structural equations $\{V_k \leftarrow f_k(Pa(k), \varepsilon_k)\}_{V_k \in V}$, quantify the causal effects shown in the causal graph G. By assuming independence among exogenous variables $\{\varepsilon_k\}_k$, the Causal Markov Condition states that $\mathrm{P}(\{V_k = v_k\}_{V_k \in V}) = \Pi_k \mathrm{P}(V_k = v_k | Pa(k) = pa(k))$. The back-door path from V_a to V_b, which is defined as the path that contains an arrow pointing into V_a, carries spurious association from V_a to V_b (Pearl, 2009).

4 METHODOLOGY

We build our causal model associated with **C**ausal **I**nvariant **Me**chanism (CIMe, *i.e.*, f_x, f_y) and **priori** about the generating process in section 4.1, followed by our identifiability result for CIMe in section 4.2, and finished by learning method to estimate CIMe in section 4.3.

4.1 *LA*TENT *C*AUSAL *I*NVARIANCE *Model*

We introduce latent variables to model the abstractions/concepts that play as causal factors that generate the observed variables (X, Y). We explicitly separate the latent variables into two parts: the S and Z that respectively denote the y-causative and y-non-causative factors, as shown by the arrow $S \to Y$ in Fig. 1. Besides, we denote the generating mapping $f_x : (S, Z) \to X, f_y : S \to Y$ as the **C**ausal **I**nvariant **Me**chanisms (CIMe) that hold across all domains. The output Y denotes the label generated by human knowledge, *e.g.*, the semantic shape, the contour to discern the object, etc. Hence, we assume the Y as the outcome/effect of these high-level abstractions (Biederman, 1987) rather than the cause (detailed comparison with $Y \to S$ is left in supplementary 7.6.1). We call the model associated with the causal graph in Fig. 1 as **La**tent **C**ausal **I**nvariance **M**odel (LaCIM), with formal definition given in Def. 4.1.

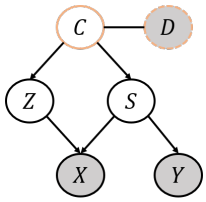


Figure 1: The DAG for LaCIM. The variables marked by white (gray) color represent the unobserved (observed) variables. Each arrow represents the causal effect from the variable it points from on the one it points to. The confounder C means the source variable of the domain, and the S, Z are the causal factors of X, Y. We observe domain index D as a substitute when C is not provided.

If X, Y denote the image and class label, the concepts (S, Z) can denote light, angle, the shape of the object to generate X following the physical mechanisms. Among these concepts, only the ones that are related to the object, *i.e.*, S (*e.g.*, shape) are causal factors of the object label, *i.e.*, Y. Following the physical or natural law, the mechanisms $S, Z \to X, S \to Y$ invariantly hold across domains. The $\mathcal{S} := \mathbb{R}^{p_s}, \mathcal{Z} := \mathbb{R}^{p_z}$ denote the space of S, Z, with $\mathrm{P}^e(S, Z)$ varies across $\mathcal{E}$ (*e.g.*, the object is more often associated with some specific types of scenes).

In addition, each domain is characterized by the value of confounder C (c^e for any environment e), which is also introduced in Ilse et al. (2020) and is here defined as the common cause or source attributes that generate latent variables in the corresponding domain. Returning to the object classification task, the C can denote the sampling strategy, and more specifically the frequency of different scenes associated with the same object (He et al., 2019). Therefore, it is domain-dependent and can intrinsically explain the distributional change in $\mathrm{P}^e(S, Z)$ (*i.e.*, sample selection bias in the object classification example) among domains. From the perspective of causality, the confounder C blocks the back-door path from Z to Y, making the Z spuriously correlated with Y. Such a correlation, which can be learned during the brutal-force data fitting process, may fail to generalize to other domains as the value of C varies among domains. Notably, we generally do not require C to be observed, in which cases we observe domain index D as a substitute (marked by dot circle in Fig. 1).

Definition 4.1 (LaCIM)**.** *The* **La**tent **C**ausal **I**nvariance **M**odel (LaCIM) *for $e \in \mathcal{E}$ is defined as a scm characterized by (i) the* ***causal graph****, i.e., the $G = (V, E)$ with $V = \{C, S, Z, X, Y\}$ and $E = \{C \to S, C \to Z, Z \to X, S \to X, S \to Y\}$; and (ii)* ***structural equations*** *with causal mechanisms $\{f_z, f_s, f_x, f_y\}$ embodying the quantitative causal information: $c \leftarrow c^e, z \leftarrow f_z(c^e, \varepsilon_z), s \leftarrow f_s(c^e, \varepsilon_s); x \leftarrow f_x(s, z, \varepsilon_x); y \leftarrow f_y(s, \varepsilon_y)$, in which $\{\varepsilon_z, \varepsilon_s, \varepsilon_x, \varepsilon_y\}$ are independent exogenous variables that induce $p_{f_z}(z|c^e), p_{f_s}(s|c^e), p_{f_x}(x|s, z), p_{f_y}(y|s)$. The CIMe f_x, f_y are assumed to be*

invariant across $\mathcal{E}$. We call the environment-dependent parts: prior and posterior of latent variables, i.e., $P^e(S,Z)$ and $P^e(S,Z|X)$ as S,Z-prior and S,Z-inference in the following.

We denote LaCIM-C and LaCIM-D as two versions of LaCIM respectively with C observed and not. For the LaCIM-C, the training data for each $\mathcal{D}^e$ turns to $\{x_i^e, y_i^e, c^e\}_{i\in[n_e]}$ with c^e additionally observed. This can be possible in some applications (*e.g.*, the C refers to age group, gender that characterizes population in image-based medical diagnosis), in which the distributional inconsistency can be "explained away" by observing C. As for the LaCIM-D with C unobserved, we use domain index D as a substitute, which is a more general setting.

For both versions, we need to identify the CIMe, accomplished by the identifiability of CIMe that enables the learning method to distinguish S from Z for prediction to avoid spurious correlation, as presented in section 4.2. A reformulated VAE is proposed to learn the CIMe practically. For generalization, note that the gap between two environments in terms of prediction given x, *i.e.*, $\left|\mathbb{E}_{p^{e_2}}[Y|X=x]-\mathbb{E}_{p^{e_1}}[Y|X=x]\right| = \int_{\mathcal{S}}\left|p^{e_2}(s|x)-p^{e_1}(s|x)\right|p_{f_y}(y|s)ds$, is mainly due to the inconsistency of S,Z-inference, *i.e.*, $p^e(s,z|x)$. Since $p^e(s,z|x)=p(s,z|x,c^e)$, it can be trained from $\mathcal{E}_{\text{train}}$, under the premise of no unobserved confounder of S,Z (corresponds to the back-door criterion in scm (Pearl, 2009)). This premise may not hold in some applications. However, we empirically find that observing partial confounders can also bring OOD generalization benefit, with alleviation of the S,Z-inference's gap. For LaCIM-D with C unobserved, we propose to inference s,z that give rise to the test sample x via maximizing the identified $p_{f_x}(x|s,z)$, as a pseudo-likelihood of x given (s,z), rather than using S,Z-inference model which is inconsistent among environments.

4.2 Identifiability of Causal Invariant Mechanisms

We present the identifiable claim about the CIMe f_x, f_y, which implicitly distinguishes the y-causative factors (*a.k.a*, S) from others (*a.k.a*, Z) for prediction, to provide a theoretical guarantee for avoiding spurious correlations. Notably, the S and Z play "asymmetric roles" in terms of generating process, as reflected in additional generating flow from S to Y. This "information intersection" property of S, *i.e.*, $f_y^{-1}(\bar{y}) = [f_x^{-1}]_S(\bar{x})$ for any $(\bar{x},\bar{y}) \in f_x(\mathcal{S},\mathcal{Z}) \times f_y(\mathcal{S})$, is exploited to disentangle S from Z. Such a disentanglement analysis, is crucial to causal prediction but lacked in existing literature (Khemakhem, Kingma and Hyvärinen, 2020; Khemakhem, Monti, Kingma and Hyvärinen, 2020; Teshima et al., 2020). Besides, our analysis extends the scope of Khemakhem, Kingma and Hyvärinen (2020) to categorical Y and general forms of $\mathrm{P}(S,Z)$ that belongs to Sobolev space, in theorem 4.4. Note that our analysis does NOT require observing the confounder C.

We assume the *Additive Noise Model* (ANM) for $f_x(s,z,\varepsilon_x) = \hat{f}_x(s,z)+\varepsilon_x$ (we replace $\hat{f}_x$ with f_x without loss of generality), which has been widely adopted to identify the causal factors (Janzing et al., 2009; Peters et al., 2014; Khemakhem, Kingma and Hyvärinen, 2020). We first narrow our interest to a family class of LaCIM denoted as $\mathcal{P}_{\text{exp}}$ such that any $p \in \mathcal{P}_{\text{exp}}$ satisfies that (i) the S,Z belong to the exponential family; and that (ii) the Y is generated from the ANM. We will show later that $\mathcal{P}_{\text{exp}}$ can approximate any general form with $\mathrm{P}(S,Z) \in W^{r,2}(\mathcal{S}\times\mathcal{Z})$ for some $r \geq 2$:

$$\mathcal{P}_{\text{exp}} = \left\{\text{LaCIM with } \; y = f_y(s)+\varepsilon_y, p^e(s,z) = p_{\mathbf{T}^z,\mathbf{\Gamma}_e^z}(z)p_{\mathbf{T}^s,\mathbf{\Gamma}_e^s}(s)\right\}$$
$$\text{with } \; p_{\mathbf{T}^t,\mathbf{\Gamma}_e^t}(t) := \prod_{i=1}^{p_t}\exp\left(\sum_{j=1}^{k_t}T_{i,j}^t(t_i)\Gamma_{e,i,j}^t + B_i(t_i) - A_{e,i}^t\right) \text{ for } t=s,z, \text{ and } e\in\mathcal{E}, \tag{1}$$

where $\{T_{i,j}^t(t_i)\}, \{\Gamma_{e,i,j}^t\}$ denote the sufficient statistics and natural parameters, $\{B_i\}, \{A_t^i\}$ denote the base measures and normalizing constants to ensure the integral of distribution equals to 1. Let $\mathbf{T}^t := \left[\mathbf{T}_1^t, ..., \mathbf{T}_{p_t}^t\right] \in \mathbb{R}^{k_t\times p_t}$ $\left(\mathbf{T}_i^t := [T_{i,1}^t, ..., T_{i,k_t}^t], \forall i \in [p_t]\right)$ and $\mathbf{\Gamma}_e^t := \left[\Gamma_{e,1}^t, ..., \Gamma_{e,p_t}^t\right] \in \mathbb{R}^{k_t\times p_t}$ $\left(\mathbf{\Gamma}_{e,i}^t := [\Gamma_{e,i,1}^t, ..., \Gamma_{e,i,k_t}^t], \forall i \in [p_t]\right)$. We define the $\sim_p$-identifiability for $\theta := \{f_x, f_y, \mathbf{T}^s, \mathbf{T}^z\}$ as:

Definition 4.2 ($\sim_p$-identifiability)**.** *We define a binary relation on the parameter space of $\mathcal{X}\times\mathcal{Y}$: $\theta \sim_p \tilde{\theta}$ if there exist two sets of permutation matrices and vectors, (M_s, a_s) and (M_z, a_z) for s and z respectively, such that for any $(x,y) \in \mathcal{X}\times\mathcal{Y}$,*

$$\mathbf{T}^s([f_x^{-1}]_{\mathcal{S}}(x)) = M_s\tilde{\mathbf{T}}^s([\tilde{f}_x^{-1}]_{\mathcal{S}}(x)) + a_s, \; \mathbf{T}^z([f_x^{-1}]_{\mathcal{Z}}(x)) = M_z\tilde{\mathbf{T}}^z([\tilde{f}_x^{-1}]_{\mathcal{Z}}(x)) + a_z,$$
$$p_{f_y}(y|[f_x^{-1}]_{\mathcal{S}}(x)) = p_{\tilde{f}_y}(y|[\tilde{f}_x^{-1}]_{\mathcal{S}}(x)).$$

We say that θ is $\sim_p$-identifiable, if for any $\tilde{\theta}$, $p_\theta^e(x,y) = p_{\tilde{\theta}}^e(x,y)\ \forall e \in \mathcal{E}_{\text{train}}$, implies $\theta \sim_p \tilde{\theta}$.

It is shown in supplement 7.2 that $\sim_p$ is an equivalence relation. The definition regarding f_x implies the separation of S and Z unless the extreme case when S can be represented by Z, *i.e.*, there exists a function $h : \mathcal{S} \to \mathcal{Z}$ such that $[f_x^{-1}]_{\mathcal{S}}(x) = h([f_x^{-1}]_{\mathcal{Z}}(x))$. This definition is inspired by but beyond the scope of unsupervised scenario considered in nonlinear ICA (Hyvärinen et al., 2019; Khemakhem, Kingma and Hyvärinen, 2020) to further distinguish of S from Z. The $p_{f_y}(y|[f_x^{-1}]_{\mathcal{S}}(x)) = p_{\tilde{f}_y}(y|[\tilde{f}_x^{-1}]_{\mathcal{S}}(x))$ further guarantees the identifiability of prediction: predict using $f_y(s)$ with s obtained from f_x. The following theorem presents the $\sim_p$-identifiability for $\mathcal{P}_{\text{exp}}$:

Theorem 4.3 ($\sim_p$-identifiability)**.** *For θ in the LaCIM $p_\theta^e(x,y) \in \mathcal{P}_{\text{exp}}$ for any $e \in \mathcal{E}_{\text{train}}$, we assume the CIMe satisfies that f_x, f_x' and f_x'' are continuous and that f_x, f_y are bijective; and that the $T_{i,j}^t$ are twice differentiable for any $t = s, z, i \in [p_t], j \in [k_t]$. Besides, we assume the exogenous variables satisfy that the characteristic functions of $\varepsilon_x, \varepsilon_y$ are almost everywhere nonzero. Under the* CONDITION *that the number of environments, i.e., $m \geq \max(p_s * k_s, p_z * k_z) + 1$ and $\left[\mathbf{\Gamma}_{e_2}^{t=s,z} - \mathbf{\Gamma}_{e_1}^{t=s,z}, ..., \mathbf{\Gamma}_{e_m}^{t=s,z} - \mathbf{\Gamma}_{e_1}^{t=s,z}\right]$ have full column rank for both $t = s$ and $t = z$, we have that the parameters $\theta := \{f_x, f_y, \mathbf{T}^s, \mathbf{T}^z\}$ are $\sim_p$ identifiable.*

The bijectivity of f_x and f_y has been widely assumed in Janzing et al. (2009); Peters et al. (2014; 2017); Khemakhem, Kingma and Hyvärinen (2020); Teshima et al. (2020) as a basic condition for identifiability. It naturally holds for f_x to be bijective since the latent components S, Z, as high-level abstractions which can be viewed as embeddings in auto-encoder (Kramer, 1991), lies in lower-dimensional space compared with input X which is supposed to have more variations, *i.e.*, $(p_s + p_z < p_x)$. For classification task with categorical Y, the f_y which generates the classification result, *i.e.*, $p(y = k|s) = [f_y]_k(s)/\left(\sum_k [f_y]_k(s)\right)$, will be shown later to be identifiable.

The diversity condition for environments can be proved to generically hold in supplementary 7.4. Such a condition, which is also assumed in Arjovsky et al. (2019), is almost necessary for the invariance to be identified from the variant ones. As noted in the formulation, a larger m would be easier to satisfy the condition, which agrees with the intuition that more environments can provide more complementary information for the identification of the invariant mechanisms.

Remark 1. *The dimensions of the ground-truth S, Z are unknown, making the check about whether m is large enough impossible. Besides, in some real applications, the training environments are passively observed and may not satisfy the condition. However, we empirically find the improvement of LaCIM in terms of both OOD prediction and interpretability, if the multiple environments provided are diverse enough. Besides, a training environment can be the mixture of many sub-environments, which motivates to splitting the data according to their source ID or clustering results (Teney et al., 2020) to obtain more environments, making the condition easier to satisfy.*

Extension to the general form of LaCIM. We generalize the identifiable result in theorem 4.3 to any LaCIM as long as its $\mathrm{P}(S, Z) \in W^{r,2}(\mathcal{S} \times \mathcal{Z})$ (for some $r \geq 2$) and categorical Y, in the following theorem. This is accomplished by showing that any such LaCIM can be approximated by a sequence of distributions in $\mathcal{P}_{\text{exp}}$, motivated by the facts in Barron and Sheu (1991) that the exponential family is dense in the set of distributions with bounded support, and in Maddison et al. (2016) that the continuous variable with multinomial logit model can be approximated by a series of distributions with *i.i.d* Gumbel noise as the temperature converges to infinity.

Theorem 4.4 (Asymptotic $\sim_p$-identifiability)**.** *Consider a LaCIM P satisfying that $p_{f_x}(x|s,z)$ and $p_{f_y}(y|s)$ are smooth w.r.t s, z and s respectively. Suppose $P(S, Z) \in W^{r,2}(\mathcal{S} \times \mathcal{Z})$ for some $r \geq 2$, we have that P is asymptotically $\sim_p$-identifiable defined as: $\forall\, \epsilon > 0$, $\exists\ \sim_p$-identifiable $\tilde{P}_\theta \in \mathcal{P}_{\text{exp}}$, s.t. $d_{\text{Pok}}(p^e(x,y), \tilde{p}_\theta^e(x,y)) < \epsilon, \forall e \in \mathcal{E}_{\text{train}}, (x,y) \in \mathcal{X} \times \mathcal{Y}$* [1].

4.3 Causal Supervised Variational Auto-Encoder

Guided by identifiability of CIMe f_x, f_y, we first provide the training method by reformulating VAE in a supervised scenario, followed by optimization over latent space for inference and test.

Training. To learn the CIMe, we reformulate the objective of VAE (Kingma and Welling, 2014) in supervised scenario. For unsupervised learning, the VAE maximizes the following **E**vidence **L**ower **B**ound (ELBO): $-\mathcal{L}_{\phi,\psi} = \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\psi(z|x)} \log p_\phi(x|z) - D_{\text{KL}}(q_\psi(z|x), p(z))\right]$, where D_{KL}

[1] The d_{Pok} denotes the Pokorov distance and $\lim_{n\to\infty} d_{\text{Pok}}(\mu_n, \mu) \to 0 \iff \mu_n \xrightarrow{d} \mu$.

is the Kullback-Leibler-Divergence. The ELBO is less and equal to $\mathbb{E}_{p(x)}[\log p_\phi(x)]$ and the equality is achieved when $q_\psi(z|x) = p_\phi(z|x)$. In our supervised scenario, we introduce the variational distribution $q^e_\psi(s, z|x, y)$ and the corresponding ELBO for any e is

$$-\mathcal{L}^e_{\phi,\psi} = \mathbb{E}_{p^e(x,y)}\left[\mathbb{E}_{q^e_\psi(s,z|x,y)} \log \frac{p^e_\phi(x, y, s, z)}{q^e_\psi(s, z|x, y)}\right]. \tag{2}$$

Since the minimizer of $\mathcal{L}^e_{\phi,\psi}$ is equivalent to the minimizer of $D_{\mathrm{KL}}(q^e_\psi(s, z|x, y), p^e_\phi(s, z|x, y))$ $+D_{\mathrm{KL}}(p^e(x, y), p^e_\phi(x, y))$, which aims at (i) learning $p^e(x, y)$ to identify CIMe; and (ii) expecting $q^e_\psi(s, z|x, y)$ to approximate the posterior $p^e_\phi(s, z|x, y)$. These induce reparameterizing $q^e_\psi(s, z|x, y)$ by $\left(q^e_\psi(s, z|x), q_\psi(y|s)\right)$, as $\frac{q^e_\psi(s,z|x)q_\psi(y|s)}{q^e_\psi(y|x)}$ to adapt to the test stage with y unobserved. According to Causal Markov Condition, we have that $p^e_\phi(x, y, s, z) = p_\phi(x|s, z)p^e_\phi(s, z)q_\psi(y|s)$ with $p_\phi(y|s)$ parameterized as $q_\psi(y|s)$ that is shared among $\mathcal{E}$. Then the $\mathcal{L}^e_{\phi,\psi}$ in Eq. (2) can be rewritten as:

$$\mathcal{L}^e_{\phi,\psi} = \mathbb{E}_{p^e(x,y)}\left[-\log q^e_\psi(y|x) - \mathbb{E}_{q^e_\psi(s,z|x)} \frac{q_\psi(y|s)}{q^e_\psi(y|x)} \log \frac{p_\phi(x|s, z)p^e_\phi(s, z)}{q^e_\psi(s, z|x)}\right], \tag{3}$$

where $q^e_\psi(y|x) = \int_{\mathcal{S}} q^e_\psi(s|x)q_\psi(y|s)ds$. The overall loss function is: $\mathcal{L}_{\phi,\psi} = \sum_{e\in\mathcal{E}_{\text{train}}} \mathcal{L}^e_{\phi,\psi}$. The training datasets $\{\mathcal{D}^e\}_{e\in\mathcal{E}_{\text{train}}}$ are applied to optimize prior model $p^e_\phi(s, z)$, inference model $q^e_\psi(s, z|x)$ and generative models $p_\phi(x|s, z), q_\psi(y|s)$ in Eq. (3). The generative models $p_\phi(x|s, z), q_\psi(y|s)$ are shared among all environments, while the $p^e_\phi(s, z), q^e_\psi(s, z|x)$ are respectively $p_\phi(s, z|c^e), q_\psi(s, z|x, c^e)$ and $p_\phi(s, z|d^e), q_\psi(s, z|x, d^e)$ for LaCIM-C and LaCIM-D.

Inference & Test. When $c^{e_{\text{test}}}$ can be acquired during test for $e_{\text{test}} \in \mathcal{E}_{\text{test}}$, we can predict y as $\arg\max_y q_\psi(y|x, c^{e'}) = \int q_\psi(s|x, c^{e'})q_\psi(y|s)ds$. Otherwise, for LaCIM-$D$ with C unobserved, we first optimize s, z via $(s^\star, z^\star) := \arg\max_{s,z} \log p_\phi(x|s, z)$ and predict y as $\arg\max_y q_\psi(y|s^\star)$. Specifically, we adopt the strategy for optimization in Schott et al. (2018) that we first sample initial points and select the one with the maximum $\log p_\phi(x|s, z)$, then we optimize for 50 iterations using Adam. The implementation details and optimization effect are shown in supplement 7.8.

5 Experiment

We evaluate LaCIM on (I) synthetic data to verify the identifiability in theorem 4.3; (II) OOD challenges: object classification with sample selection bias (Non-I.I.D. Image dataset with Contexts (NICO)); Hand-Writing Recognition with confounding bias (Colored MNIST (CMNIST)); prediction of Alzheimer's Disease (Alzheimer's Disease Neuroimaging Initiative (ADNI `www.loni.ucla.edu/ADNI`); (III) Robustness on detecting images with small perturbation (FaceForensics++).

5.1 Simulation

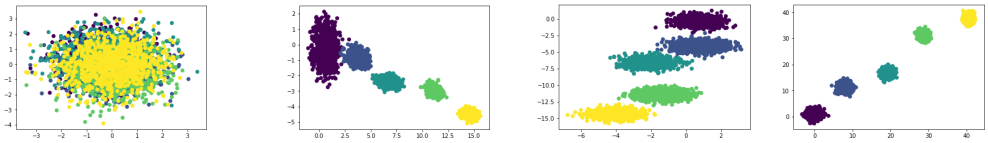
To verify the identifiability claim and effectiveness of our learning method, we implement LaCIM on synthetic data. The data generating process is provided in Supplement 7.7. The domain index $D \in \mathbb{R}^m$ is denoted as a one-hot encoded vector with $m = 5$. As verification of the utility of training on multiple domains, we also conduct LaCIM by pooling data from all m domains together, namely pool-LaCIM for comparison. We randomly generate 5 datasets and run 20 times for each. We computed the metric mean correlation coefficient (MCC) adopted in Khemakhem, Kingma and Hyvärinen (2020) to measure the goodness of identifiability under permutation by introducing cost optimization to assign each learned component to the source component. This measurement is aligned with the goal of $\sim_p$-identifiability, which allows us to distinguish S from Z. Table 5.1 shows that our LaCIM-D achieves comparable identifiable results than LaCIM-C regarding the CIMe relating to S, Z under permutation, by means of multiple diverse experiments. As an illustration of learning effect, we visualize the learned Z in Fig. 7.7, with S left in supplement 7.7 due to space limit.

5.2 Real-world OOD Challenge

We present our LaCIM's results on three OOD tasks, with different environments associated with different values of C. We implement both versions of LaCIM, *i.e.*, LaCIM-C and LaCIM-D, with

Table 1: MCC of identified latent variables. Average over 20 times for each data.

	Data #1		Data #2		Data #3		Data #4		Data #5		Average	
	Z	S	Z	S	Z	S	Z	S	Z	S	Z	S
pool-LaCIM	0.28	0.58	0.38	0.66	0.34	0.77	0.34	0.79	0.36	0.75	0.34	0.71
LaCIM-C (**Ours**)	**0.81**	**0.86**	**0.81**	**0.87**	**0.85**	**0.87**	0.73	0.78	**0.86**	**0.87**	**0.82**	**0.85**
LaCIM-D (**Ours**)	0.64	0.82	0.75	0.80	0.76	0.83	**0.79**	**0.90**	0.75	0.85	0.74	0.84



(a) $p_{\text{pool-LaCIM}}(z|x,y)$ (b) $p_{\text{LaCIM}}(z|x,y,c)$ (c) $p_{\text{LaCIM}}(z|x,y,d)$ (d) $p_{\theta^\star}(z|c)$

Figure 2: Visualization of Z. From left to right are: estimated posterior by pool-LaCIM, LaCIM-C, LaCIM-D and the ground-truth.

task-dependent definition of C. In CMNIST, the C (digit color) is a fully observed confounder, and LaCIM-C in this case is the ceiling of LaCIM-D under the same implementation. In NICO and ADNI, the LaCIM-D even outperform LaCIM-C, due to the existence of unobserved confounders.

Dataset. We describe the datasets as follows (details about data and implementation are respectively left in supplement 7.10, 7.9, 7.11):

NICO: we evaluate the cat/dog classification in "Animal" dataset in NICO, a benchmark for non-i.i.d problem in He et al. (2019). Each animal is associated with "grass","snow" contexts with different proportions, denoted as $C \in \mathbb{R}^4$ (cat,dog in grass,snow). We consider two settings: $m = 8, 14$.

CMNIST: We relabel the digits 0-4 and 5-9 as $y = 0$ and $y = 1$, based on MNIST. Then we color p^e of images with $y = 0$ and $1 - p^e$ of images with $y = 1$ as green and color others as red. We set $m = 2$ with $p^{e_1} = 0.9, p^{e_2} = 0.8$. The c^e is p^e to describe the intensity of confounding bias caused by color. We do not flip y with 25% like Arjovsky et al. (2019) [2], since doing so will cause the digit correlated rather than causally related to the label, which is beyond our scope.

ADNI. The data are obtained from the ADNI databaset, the $\mathcal{Y} := \{0, 1, 2\}$ with 0 denoting AD, 1 denoting Mild Cognitive Impairment (MCI) and 2 denoting Normal Control (NC). The X is Magnetic resonance imaging (sMRI). We set $m = 2$. We consider two types of C: Age and TAU (a biomarker Humpel and Hochstrasser (2011)).

Compared Baselines. We compare with (i) Cross-Entropy (CE) from $X \to Y$ (CE $X \to Y$), (ii) domain-adversarial neural network (DANN) for domain adaptation Ganin et al. (2016), (iii) Maximum Mean Discrepancy with Adversarial Auto-Encoder (MMD-AAE) for domain generalization Li et al. (2018), (iv) Invariant Risk Mnimization (IRM) Arjovsky et al. (2019), (v) CE $(X, C) \to Y$, (vi) VAE with causal graph $C \to V \to \{X, Y\}$ with V mixing S, Z and we call it sVAE for simplicity.

Results & Discussions. We report accuracy over three runs for each method. As shown in Tab. 2 [3], our LaCIM-D performs comparable and better than others on all applications. Especially, the advantage over invariant learning method (IRM) and CE $(X, C) \to Y$ which also takes C into prediction can be contributed to the identification of true causal mechanisms. Further, the improvement over sVAE is benefited from our separation of y-causative factors (*a.k.a*, S) from others to avoid spurious correlation. Besides, as shown from results on NICO, a larger m (with the total number of samples n fixed) can bring further benefit, which may due to the easier satisfaction of the diversity condition in theorem 4.3. One thing worth particular mention is that on NICO and ADNI (when C denotes TAU), our LaCIM-D performs comparable and even better than LaCIM-C, due to the existence of unobserved confounders. For example, each C only contains one attribute each time in ADNI. For completeness, we conduct experiments with fully observed confounders in supplement 7.11.

[2] We also conduct this experiment with flipping y in supplementary 7.9.

[3] On NICO, we implement ConvNet with Batch Balancing as a specifically benchmark in He et al. (2019). The results are 60 ± 1 on $m = 8$ and 62.33 ± 3.06 on $m = 14$.

Table 2: Accuracy (%) of OOD prediction. Average over three runs.

Method \ Dataset	NICO		CMNIST	ADNI ($m = 2$)	
	$m = 8$	$m = 14$	$m = 2$	C : Age	C : TAU
CE $X \to Y$	58.67 ± 5.69	59.33 ± 4.04	97.82 ± 0.20	63.06 ± 2.26	64.58 ± 0.90
DANN	59.00 ± 2.65	58.67 ± 2.08	97.31 ± 0.30	60.84 ± 1.83	64.58 ± 0.90
MMD-AAE	60.33 ± 2.31	65.33 ± 5.69	98.09 ± 0.21	59.79 ± 1.73	65.62 ± 0.00
IRM	60.67 ± 2.08	63.00 ± 1.00	98.08 ± 0.25	63.49 ± 1.59	65.10 ± 0.90
CE $X, C \to Y$	57.33 ± 6.03	64.00 ± 1.00	97.17 ± 0.74	62.43 ± 0.92	65.62 ± 0.00
sVAE	59.67 ± 3.79	64.33 ± 0.58	97.89 ± 0.61	63.67 ± 1.87	66.67 ± 0.91
LaCIM-C (**Ours**)	62.00 ± 1.73	68.00 ± 2.64	$\mathbf{98.81 \pm 0.14}$	$\mathbf{65.08 \pm 1.59}$	66.14 ± 0.91
LaCIM-D (**Ours**)	$\mathbf{62.67 \pm 0.58}$	$\mathbf{68.67 \pm 2.64}$	98.78 ± 0.20	64.44 ± 0.96	$\mathbf{68.23 \pm 0.90}$

Interpretability. We visualize learned S as side proof of interpretability. As shown in Fig. 5.2, LaCIM (the 4th column) can identify more explainable semantic features, which verifies the identifiability and effectiveness of the learning method. Supplement 7.10 provides more results.

(a) Cat on grass (b) Cat on snow

(c) Dog on grass (d) Dog on snow

Figure 3: Visualization via gradient Simonyan et al. (2013). From the left to right: original image, CE $X \to Y$, CE $(X, C) \to Y$ and LaCIM-C.

5.3 Robustness on Security

We consider the DeepFake-related security problem, in which the goal is detecting small perturbed images that can spread fake news. The Rossler et al. (2019) provides FaceForensics++ dataset from 1000 Youtube videos for training and 1,000 benchmark images from other sources (OOD) for testing. There are in total 4 ways for manipulation: (i) replacement by Deepfake, FaceSwap; and (ii) reenactment by FacetoFace, NeuralTextures. Inspired by Teney et al. (2020), we split the train data into $m = 2$ environments according to video ID. The considerable result [4] in Tab. 5.3 verifies potential value on security.

Table 3: Accuracy (%) of robustness on FaceForensics++. Average over three runs.

CE $X \to Y$	IRM	LaCIM-D (**Ours**)
82.8 ± 0.99	83.4 ± 0.59	$\mathbf{84.47 \pm 0.90}$

6 Conclusions & Discussions

We incorporate the causal structure as prior knowledge in proposed LaCIM, by introducing: (i) latent variables and explicitly separate them into y-causative factors (*a.k.a*, S) and others (*a.k.a*, Z) which are spuriously correlated with the output; (ii) the source variable C that explains the distributional inconsistency among domains. When the environments are diverse and much enough, we can successfully identify the causal invariant mechanisms, and also y-causative factors for prediction without a mix of others. Our LaCIM shows potential value regarding robustness to OOD tasks with confounding bias, selection bias and others such as healthcare and security. A possible drawback of our model lies in our requirement of the number of environments (which may be not satisfied in some scenarios) for identifiability, and the relaxation of which is left in the future work.

[4] Our goal is verifying the robustness of LaCIM, rather than winning the leader-board.

REFERENCES

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

Barron, A. R. and Sheu, C.-H. (1991), 'Approximation of density functions by sequences of exponential families', *The Annals of Statistics* pp. 1347–1369.

Bellot, A. and van der Schaar, M. (2020), 'Generalization and invariances in the presence of unobserved confounding', *arXiv preprint arXiv:2007.10653* .

Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. (2007), Analysis of representations for domain adaptation, *in* 'Advances in neural information processing systems', pp. 137–144.

Ben-Tal, A., El Ghaoui, L. and Nemirovski, A. (2009), *Robust optimization*, Vol. 28, Princeton University Press.

Bengio, Y. (2017), 'The consciousness prior', *arXiv preprint arXiv:1709.08568* .

Bengio, Y., Courville, A. and Vincent, P. (2013), 'Representation learning: A review and new perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8), 1798–1828.

Biederman, I. (1987), 'Recognition-by-components: a theory of human image understanding.', *Psychological review* **94**(2), 115.

Bühlmann, P. (2018), 'Invariance, causality and robustness', *arXiv preprint arXiv:1812.08233* .

Döbler, C. et al. (2015), 'Stein's method of exchangeable pairs for the beta distribution and generalizations', *Electronic Journal of Probability* **20**.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V. (2016), 'Domain-adversarial training of neural networks', *Journal of Machine Learning Research* **17**, 1–35.

Gatys, L. A., Ecker, A. S. and Bethge, M. (2015), 'A neural algorithm of artistic style', *arXiv preprint arXiv:1508.06576* .

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C. and Schölkopf, B. (2016), Domain adaptation with conditional transferable components, *in* 'International Conference on Machine Learning', pp. 2839–2848.

Goodfellow, I. J., Shlens, J. and Szegedy, C. (2014), 'Explaining and harnessing adversarial examples', *arXiv preprint arXiv:1412.6572* .

Grother, P. J. (1995), 'Nist special database 19 handprinted forms and characters database', *National Institute of Standards and Technology* .

Guerreiro, R. and Bras, J. (2015), 'The age factor in Alzheimer's disease', *Genome medicine* **7**(1), 106.

He, Y., Shen, Z. and Cui, P. (2019), 'Towards non-i.i.d. image classification: A dataset and baselines', *arXiv preprint arXiv:1906.02899* .

Heinze-Deml, C. and Meinshausen, N. (2017), 'Conditional variance penalties and domain shift robustness', *arXiv preprint arXiv:1710.11469* .

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B. and Smola, A. J. (2007), Correcting sample selection bias by unlabeled data, *in* 'Advances in Neural Information Processing Systems', pp. 601–608.

Humpel, C. and Hochstrasser, T. (2011), 'Cerebrospinal fluid and blood biomarkers in Alzheimer's disease', *World journal of psychiatry* **1**(1), 8.

Hyvarinen, A. and Morioka, H. (2016), Unsupervised feature extraction by time-contrastive learning and nonlinear ica, *in* 'Advances in Neural Information Processing Systems', pp. 3765–3773.

Hyvärinen, A. and Pajunen, P. (1999), 'Nonlinear independent component analysis: Existence and uniqueness results', *Neural Networks* **12**(3), 429–439.

Hyvärinen, A., Sasaki, H. and Turner, R. (2019), Nonlinear ICA using auxiliary variables and generalized contrastive learning, *in* 'The 22nd International Conference on Artificial Intelligence and Statistics', pp. 859–868.

Ilse, M., Tomczak, J. M. and Forré, P. (2020), 'Designing data augmentation for simulating interventions', *arXiv preprint arXiv:2005.01856* .

Ilse, M., Tomczak, J. M., Louizos, C. and Welling, M. (2019), 'DIVA: Domain invariant variational autoencoders', *arXiv preprint arXiv:1905.10427* .

Janzing, D., Peters, J., Mooij, J. and Schölkopf, B. (2009), Identifying confounders using additive noise models, *in* 'Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)', AUAI Press, pp. 249–257.

Johansson, F. D., Sontag, D. and Ranganath, R. (2019), Support and invertibility in domain-invariant representations, *in* 'The 22nd International Conference on Artificial Intelligence and Statistics', pp. 527–536.

Kang, G., Dong, X., Zheng, L. and Yang, Y. (2017), 'Patchshuffle regularization', *arXiv preprint arXiv:1707.07103* .

Khemakhem, I., Kingma, D. P. and Hyvärinen, A. (2020), Variational autoencoders and nonlinear ICA: A unifying framework, *in* 'Proceedings of the 23th International Conference on Artificial Intelligence and Statistics (AISTATS-23)', Vol. 108, AISTATS Committee, PMLR, Palermo, Italy.

Khemakhem, I., Monti, R. P., Kingma, D. P. and Hyvärinen, A. (2020), 'Ice-beem: Identifiable conditional energy-based deep models', *arXiv preprint arXiv:2002.11537* .

Kingma, D. P. and Welling, M. (2014), Auto-encoding variational Bayes, *in* 'Proceedings of the International Conference on Learning Representations (ICLR 2014)', ICLR Committee, Banff, Canada.

Kramer, M. A. (1991), 'Nonlinear principal component analysis using autoassociative neural networks', *AIChE journal* **37**(2), 233–243.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Priol, R. L. and Courville, A. (2020), 'Out-of-distribution generalization via risk extrapolation (rex)', *arXiv preprint arXiv:2003.00688* .

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', pp. 1617–1626.

Li, D., Yang, Y., Song, Y.-Z. and Hospedales, T. M. (2017), 'Learning to generalize: Meta-learning for domain generalization', *arXiv preprint arXiv:1710.03463* .

Li, H., Jialin Pan, S., Wang, S. and Kot, A. C. (2018), Domain generalization with adversarial feature learning, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5400–5409.

Maddison, C. J., Mnih, A. and Teh, Y. W. (2016), 'The concrete distribution: A continuous relaxation of discrete random variables', *arXiv preprint arXiv:1611.00712* .

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2017), 'Towards deep learning models resistant to adversarial attacks', *arXiv preprint arXiv:1706.06083* .

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P. and Mooij, J. M. (2018), Domain adaptation by using causal inference to predict invariant conditional distributions, *in* 'Advances in Neural Information Processing Systems', pp. 10846–10856.

Marcos, D., Volpi, M. and Tuia, D. (2016), Learning rotation invariant convolutional filters for texture classification, *in* '2016 23rd International Conference on Pattern Recognition (ICPR)', IEEE, pp. 2012–2017.

Mortimer, J. A. (1997), 'Brain reserve and the clinical expression of Alzheimer's disease.', *Geriatrics (Basel, Switzerland)* **52**, S50–3.

Muandet, K., Balduzzi, D. and Schölkopf, B. (2013), Domain generalization via invariant feature representation, *in* 'International Conference on Machine Learning', pp. 10–18.

Pan, S. J., Tsang, I. W., Kwok, J. T. and Yang, Q. (2010), 'Domain adaptation via transfer component analysis', *IEEE Transactions on Neural Networks* **22**(2), 199–210.

Pearl, J. (2009), *Causality*, Cambridge university press.

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.

Peters, J., Janzing, D. and Schölkopf, B. (2017), *Elements of causal inference: foundations and learning algorithms*, MIT press.

Peters, J., Mooij, J. M., Janzing, D. and Schölkopf, B. (2014), 'Causal discovery with continuous additive noise models', *Journal of Machine Learning Research* **15**(1), 2009–2053.

Rojas-Carulla, M., Schölkopf, B., Turner, R. and Peters, J. (2018), 'Invariant models for causal transfer learning', *The Journal of Machine Learning Research* **19**(1), 1309–1342.

Romeijn, J.-W. and Williamson, J. (2018), 'Intervention and identifiability in latent variable modelling', *Minds and machines* **28**(2), 243–264.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M. (2019), Faceforensics++: Learning to detect manipulated facial images, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1–11.

Sagawa, S., Koh, P. W., Hashimoto, T. B. and Liang, P. (2019), 'Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization', *arXiv preprint arXiv:1911.08731* .

Schölkopf, B. (2019), 'Causality for machine learning', *arXiv preprint arXiv:1911.10500* .

Schölkopf, B., Janzing, D., Peters, J. and Zhang, K. (2011), 'Robust learning via cause-effect models', *arXiv preprint arXiv:1112.2738* .

Schott, L., Rauber, J., Bethge, M. and Brendel, W. (2018), 'Towards the first adversarially robust neural network model on mnist', *arXiv preprint arXiv:1805.09190* .

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P. and Sarawagi, S. (2018), Generalizing across domains via cross-gradient training, *in* 'Proceedings of the International Conference on Learning Representations (ICLR 2018)'.

Shorten, C. and Khoshgoftaar, T. M. (2019), 'A survey on image data augmentation for deep learning', *Journal of Big Data* **6**(1), 60.

Simonyan, K., Vedaldi, A. and Zisserman, A. (2013), 'Deep inside convolutional networks: Visualising image classification models and saliency maps', *arXiv preprint arXiv:1312.6034* .

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P. and Kawanabe, M. (2008), 'Direct importance estimation for covariate shift adaptation', *Annals of the Institute of Statistical Mathematics* **60**(4), 699–746.

Suter, R., Miladinovic, D., Schölkopf, B. and Bauer, S. (2019), Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness, *in* 'International Conference on Machine Learning', pp. 6056–6065.

Tan, M. and Le, Q. V. (2019), 'Efficientnet: Rethinking model scaling for convolutional neural networks', *ArXiv* **abs/1905.11946**.

Taylor, L. and Nitschke, G. (2017), 'Improving deep learning using generic data augmentation', *arXiv preprint arXiv:1708.06020* .

Teney, D., Abbasnejad, E. and Hengel, A. v. d. (2020), 'Unshuffling data for improved generalization', *arXiv preprint arXiv:2002.11894* .

Teshima, T., Sato, I. and Sugiyama, M. (2020), 'Few-shot domain adaptation by causal mechanism transfer', *arXiv preprint arXiv:2002.03497* .

Vina, J. and Lloret, A. (2010), 'Why women have more Alzheimer's disease than men: gender and mitochondrial toxicity of amyloid-β peptide', *Journal of Alzheimer's disease* **20**(s2), S527–S533.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D. and Brostow, G. J. (2017), Harmonic networks: Deep translation and rotation equivariance, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5028–5037.

Xie, C., Chen, F., Liu, Y. and Li, Z. (2020), 'Risk variance penalization: From distributional robustness to causality', *arXiv preprint arXiv:2006.07544* .

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2016), 'Understanding deep learning requires rethinking generalization', *arXiv preprint arXiv:1611.03530* .

Zhang, K., Schölkopf, B., Muandet, K. and Wang, Z. (2013), Domain adaptation under target and conditional shift, *in* 'International Conference on Machine Learning', pp. 819–827.

Zhao, H., Combes, R. T. d., Zhang, K. and Gordon, G. J. (2019), 'On learning invariant representation for domain adaptation', *arXiv preprint arXiv:1901.09453* .