Integrating Markov Blanket Discovery into Causal Representation Learning for Domain Generalization

Naiyu Yin[1], Hanjing Wang[1], Yue Yu[2], Tian Gao[3], Amit Dhurandhar[3], and Qiang Ji[1]

[1] Rensselaer Polytechnic Institute, Troy NY 12180, USA
{yinn2, wangh36, jiq}@rpi.edu
[2] Lehigh University, Bethlehem PA 18015, USA
{yuy214}@lehigh.edu
[3] IBM Research, Yorktown Heights NJ 10598, USA
{tgao, adhuran}@us.ibm.com

Abstract. Identifying low-dimensional, semantic latent causal representations for high-dimensional data has become a dynamic field in computer vision and machine learning. Causal domain generalization methods aim to identify latent causal variables that generate input data and build invariant causal mechanisms for prediction tasks, thereby improving out-of-distribution (OOD) prediction performance. However, there is no consensus on the best approach for selecting causal variables for prediction. Existing methods typically choose causal or anti-causal variables, excluding other invariant, discriminative features. In this paper, we propose using Markov Blanket features due to their property of being the minimal set that possesses the maximum mutual information with the target. To achieve this, we establish a Causal Markov Blanket Representation Learning (CMBRL) framework, which allows for Markov Blanket discovery in the latent space. We then construct an invariant prediction mechanism using the identified Markov Blanket features, making it suitable for predictions across domains. Compared to state-of-the-art domain generalization methods, our approach exhibits robustness and adaptability under distribution shifts.

Keywords: Causal Representation Learning · Markov Blanket Discovery · Domain Generalization

1 Introduction

Modern deep-learning models have significantly improved performance across various machine learning and computer vision applications in recent decades. However, they are also known for their limitations, including poor generalization, a lack of interpretability, and fairness concerns. [35] pinpointed the primary cause of the poor generalization of modern deep-learning models stems from the spurious correlations between irrelevant features and the prediction

task through theoretical analysis and empirical evidence. These spurious correlations are caused by data biases and vary when distribution shifts, rendering the deep-learning models unreliable for the prediction of data from unseen domains. Hence, causal domain generalization methods propose to formulate prediction tasks from a causal perspective and exploit the robust, domain-invariant causality instead of correlations to improve the Out-of-distribution performance.

Recent research in causal domain generalization [22, 31–34, 39, 41, 52] utilizes structural causal models (SCMs) to delve into the fundamental mechanisms underlying data generation. SCMs effectively capture intrinsic, stable, and interpretable causal relationships within a data distribution. These studies leverage SCMs to extract latent representations that are causally related to the target variable and subsequently develop invariant prediction mechanisms based on these representations. Key differences among existing methods stem from their choices of causally related features and their approaches to learning or selecting these features. However, there is no consensus on the optimal types of causally related features to be learned or selected. While many approaches [22, 32–34, 52] focus on identifying parent variables (causal features) relative to the target, [31] contends that anti-causal features (child variables) may offer greater robustness and predictive power, particularly in vision tasks. This paper theoretically examines the best causally related variables for prediction tasks and proposes an effective, potentially efficient method for identifying these variables.

According to [12], the Markov Blanket (MB) of a variable is theoretically proven to be optimal for prediction tasks, as it is the minimal set that contains the maximum information about the target. Consequently, MB features have been extensively used for feature selection across various prediction tasks [4, 38, 47]. Recent research [55] explores the use of MB features in the latent space for domain generalization. We find that, within the framework of our proposed SCM, the influence of spurious features is effectively mitigated by conditioning on the MB features. Specifically, the MB set can block all paths between the domain variable U and the target Y, as illustrated in Figure 1. This characteristic makes MB features theoretically optimal for domain generalization prediction tasks.

We propose a framework for selecting Markov Blanket features from latent high-level variables that generate high-dimensional data. We refer to the chosen MB latent variables as Causal Markov Blanket (CMB) representations, which include the parents, children, and spouses of the target variable in the latent space. To obtain the CMB representations, we first obtain a set of identifiable latent variables using existing identifiable variational autoencoder framework. Then we propose an efficient MB discovery approach to identified the CMB representations collectively. A prediction model built using CMB representations is theoretically guaranteed to be both invariant and informative for predicting the target variable.

The **main contributions** of this paper are as follows: 1) We introduce a novel and general SCM to elucidate the data generation mechanisms underlying prediction tasks. 2) We establish a framework for discovering Markov Blanket features in the latent space. 3) We propose a three-phase algorithm for do-

main generalization prediction. We validate the effectiveness of our algorithm on benchmark datasets with distribution shifts, demonstrating improved generalization performance in the presence of these shifts.

2 Related Works

In this section, we will review previous studies focusing on two key areas: Markov Blanket discovery and domain generalization (DG) prediction.

Markov Blanket Discovery and Feature Selection. The Markov Blanket set of a variable consists of its immediate neighbors (parents, children, and spouses) in a causal graph. Popular MB discovery methods, including KS [21], HITON-MB [5], IPCMB [10], IAMB [48], and STMB [11], identify the Markov Blanket set through a series of independence tests, making MB discovery performance dependent on the accuracy of these tests. In terms of applications, multiple works [4, 12, 38, 47] have utilized MB discovery in the input space as a feature selection strategy, employing MB features for prediction. However, directly applying these methods to high-dimensional data, such as images, videos, or texts, is computationally expensive. Moreover, as noted by [32], causal and spurious features may not be disentangled in the input space since each dimension of the input data could be influenced by both types of features.

Domain Generalization Approaches. We aim to employ Causal Markov Blanket representations to enhance domain generalization performance. To this end, we compare our method both theoretically and empirically with existing domain generalization approaches. Specifically, in a domain generalization setting, we use data from one or multiple source domains to make predictions on unseen target domains. Recent domain adaptation works [22, 26, 46] that require target domain data and intervention targets during training are beyond the scope of this paper. Popular domain generalization approaches include disentangled representation learning [19, 28, 29, 42], data augmentation [33], "mix-up" strategies [14, 56, 57], adversarial training [50, 51], and meta-learning [24, 43]. These methods typically do not involve causality.

Causal Domain Generalization Approaches. Compared to non-causal DG approaches, causal approaches make assumptions about data generation via SCMs and derive their algorithms accordingly. Generally, causal approaches can be categorized by their use of intervention. Methods without interventions include stable representation learning [7, 8, 17, 18] and invariant feature learning [1–3, 6, 23, 40]. Stable representation learning methods acquire causal or anti-causal features through strategies like covariate balancing or by using SCM as a form of regularization. Invariant feature learning methods aim to identify domain-invariant causal features from multi-environmental data based on certain invariance criteria derived from causal theorems. One popular invariant feature learning approach is invariant risk minimization (IRM), which identifies invariant predictors corresponding to the parent variables of the target variable. In

particular, [55] formulates an invariant causal Markov Blanket representation learning framework that is trained with causal constraints derived from the proposed SCM. Recently, a line of work [19, 20, 32] employs identifiable variational autoencoders (iVAEs) to obtain a set of latent variables with various degrees of identifiable guarantees, using the parent variables within this set for domain generalization prediction. We follow this approach and propose to learn the MB set of the target variable from the identifiable latent representations. Methods with interventions include robust feature learning through data augmentation [33] and interventional inference-guided methods [27, 34, 52]. These methods primarily address latent confounding issues between input data and the target. While our method generalizes across domains with different biases, it cannot yet effectively handle spurious correlations resulting from latent confounders. Extending our algorithm to handle latent confounders is an interesting future direction but is beyond the scope of this paper.

3 Causal Analysis for Prediction Tasks

3.1 Preliminary on Markov Blanket Features

We outline the definition of Markov Blanket in **Definition 1**.

Definition 1. *([37]) A Markov Blanket of a target variable T within the variable set $\boldsymbol{V}$, $\boldsymbol{MB}_T$, is the minimal set of nodes conditioned on which all other nodes are independent of T, denoted as $O \perp\!\!\!\perp T|\boldsymbol{MB}_T, \forall O \in \{\boldsymbol{V} \backslash T \backslash \boldsymbol{MB}_T\}$.*

The $\mathbf{MB}_T$ consists of **the parent, child, and spouse variables** of the target variable T. According to the **Theorem 1-3** in [12], the MB features of the target possess the following properties: they constitute **the minimal set** of features holding **the maximal information** about the target variable, and they ensure the **least Bayes errors** when predicting the target.

Unlike prior works that use MB discovery for feature selection in the input space, we focus on MB selection in the latent space, guided by a structural causal model that we introduce for modeling the data generation process. To distinguish between them, we denote the MB set in the latent representation space as Causal Markov Blanket (CMB) representations. CMB representations are crucial for identifying the informative variables for predicting targets with higher accuracy despite distribution shifts.

3.2 Assumptions on Structural Causal Model

We formulate the prediction task under distribution shifts from a causal perspective. Using a structural causal model, we discern the underlying causal mechanisms of the data generation process, as illustrated in Figure 1.

$\boldsymbol{X} \in \mathbb{R}^d$ represents the inputs and Y the prediction target. We denote $\boldsymbol{Z} = \{Z_1, Z_2, \cdots, Z_N\} \in \mathbb{R}^N$ as the latent factors for generating input $\boldsymbol{X}$. To better account for the distribution shifts, we introduce a domain variable U to encode domain-specific information[4]. Domain variable U is the common cause for some latent factors in $\boldsymbol{Z}$ and hence can change the appearance of input $\boldsymbol{X}$. Judged by their relations to the Y, $\boldsymbol{Z}$ can be further categorized into four types: parent variables $\boldsymbol{Z}_p = \{Z_{p_1}, Z_{p_2}, \cdots\}$, child variables $\boldsymbol{Z}_c = \{Z_{c_1}, Z_{c_2}, \cdots\}$, spouse variables $\boldsymbol{Z}_s = \{Z_{s_1}, Z_{s_2}, \cdots\}$, and spurious variables $\boldsymbol{Z}_o = \{Z_{o_1}, Z_{o_2}, \cdots\}$. $\boldsymbol{Z}_p, \boldsymbol{Z}_c$ and $\boldsymbol{Z}_s$ are the direct causes, direct effects, and the causes of the direct effects. They collectively form the CMB representations $\boldsymbol{Z}_{cmb}$. $\boldsymbol{Z}_o$ are the variables that are spuriously correlated to target Y via $\boldsymbol{Z}_{cmb}$ variables or domain variable U. To generalize our SCM, we allow for arbitrary causal relations within $\boldsymbol{Z}$ as long as the causal graph over $\{\boldsymbol{X}, Y, \boldsymbol{Z}, U\}$ is a directed acyclic graph (DAG). We summarize the detailed assumptions in **Assumption 1** of Appendix A.1. We show that our assumptions is practical and generally hold in real-world applications in Appendix A.2. Moreover, the SCM is Figure 1, combined with the assumptions, covers most scenarios from prior works [3, 32] and hence is a flexible model for performing causal analysis on prediction tasks. The semantic meanings of $\boldsymbol{Z}_o$ and $\boldsymbol{Z}_{cmb}$ are context-dependent and vary with the data. For instance, in CMNIST, $\boldsymbol{Z}_o$ represents background information like color, whereas $\boldsymbol{Z}_{cmb}$ denotes foreground information such as digit shape.

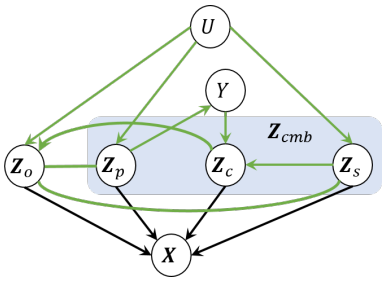

Fig. 1: Proposed SCM over $\{\boldsymbol{X}, U, \boldsymbol{Z}, Y\}$. Undirected links between $\boldsymbol{Z}_o, \boldsymbol{Z}_p$ and $\boldsymbol{Z}_s$ indicate that both edge directions are permissible without violating the DAG constraint.

3.3 Invariant Predictive Mechanism

The conditional distribution $p(Y|\boldsymbol{X})$, typically captured by a traditional classifier, is influenced by the domain variable U, meaning that $U \not\perp\!\!\!\perp Y|\boldsymbol{X}$. Consequently, the domain-variant distribution $p(Y|\boldsymbol{X})$ is unsuitable for prediction in unseen domains. Prior causal approaches such as [22, 33, 34] and other content-style approaches make strong assumptions, assuming there is only a single type of causal feature (usually parent variables $\boldsymbol{Z}_p$) within the SCM. Due to the simplicity of their SCMs, these methods fail when encountering scenarios where other types of causal features exist, as their selection of causal features cannot block all paths from U to Y. For example, if the SCM only considers the existence of parent variables $\boldsymbol{Z}_p$ while $\boldsymbol{Z}_c$ also exists in practice, the domain variable U can reach Y via $U \to \boldsymbol{Z}_o \to \boldsymbol{X} \to \boldsymbol{Z}_c \leftarrow Y$ when $\boldsymbol{X}$ is given.

[4] This is a standard setting in previous works [32, 39]

We advocate for using the Causal Markov Blanket features, denoted as $\boldsymbol{Z}_{cmb} = \{\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s\}$, to predict the target in unseen domains. From Figure 1, we infer that $Y \perp\!\!\!\perp U | \{\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s\}$, implying that $p(Y|\boldsymbol{Z}_{cmb})$ remains invariant across domains. Furthermore, we assert that the CMB features constitute the minimal sufficient set required for achieving invariant predictions across varying domains, as adding or removing variables from this set may affect its invariance. We demonstrate how to identify the CMB latent variables $\boldsymbol{Z}_{cmb}$ from the given data distribution and construct the invariant prediction mechanism $p(Y|\boldsymbol{Z}_{cmb})$ for out-of-distribution prediction in Section 4.

4 Causal Markov Blanket Representation Learning

We refer to our proposed method as CMBRL, representing **C**ausal **M**arkov **B**lanket **R**epresentation **L**earning. The training and inference procedures are introduced in Section 4.1 and Section 4.2, respectively.

4.1 Training Procedure

We illustrate the training procedure of the proposed algorithm in Figure 2. We

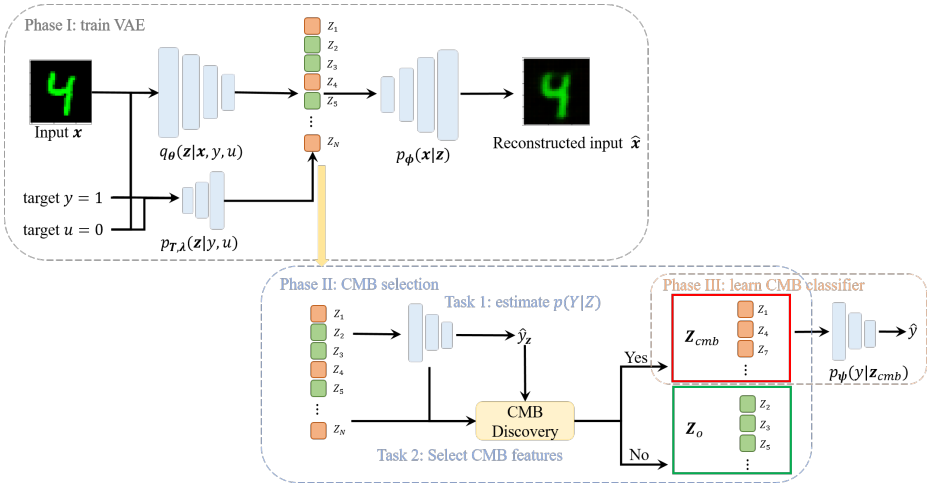

Fig. 2: The illustration of the three-phase training procedure of CMBRL.

propose a three-phase training procedure. Initially, we utilize a VAE framework to learn a set of latent variables, which are proven to be component-wise identifiable. Subsequently, we conduct a CMB search using our novel mutual information quantification method to select the latent CMB representations. Finally, we construct a predictor using the selected CMB representations.

Phase I: Identifiable Latent Variables Learning. We aim to obtain a set of latent variables $\boldsymbol{Z}$. To do so, we employ an existing (identifiable) variational autoencoder (VAE) framework [32]. We posit a prior distribution on $\boldsymbol{Z}$ that is consistent with our **Assumption 1(b)** and belongs to a general exponential family, i.e.,

$$p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{Z}|Y,U) = \frac{\mathcal{Q}(\boldsymbol{Z})}{\mathcal{C}(Y,U)} \exp[\boldsymbol{T}(\boldsymbol{Z})^T \boldsymbol{\lambda}(Y,U)]$$

where $\mathcal{Q}$ is the base measure. $\mathcal{C}$ is the normalizing constant. $\boldsymbol{\lambda}$ is the arbitrary function. $\boldsymbol{T}$ is the sufficient statistics.[5] To manage the increasing optimization challenges caused by additional parameters in the prior, we train the VAE framework using a standard protocol from prior work. [16,49]. As described in Eq. (1), the training objective comprises an ELBO loss $\mathcal{L}_{\text{ELBO}}$ and a score matching loss $\mathcal{L}_{\text{SM}}$. The ELBO loss $\mathcal{L}_{\text{ELBO}}$ optimizes over encoder and decoder parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$, while the score matching loss $\mathcal{L}_{\text{SM}}$ minimizes over prior parameters $(\boldsymbol{T}, \boldsymbol{\lambda})$. The parameters $\boldsymbol{T}, \boldsymbol{\lambda}$ are constants in $\mathcal{L}_{\text{ELBO}}$, and $\boldsymbol{\theta}, \boldsymbol{\phi}$ are constants in $\mathcal{L}_{\text{SM}}$.

$$\mathcal{L}_{\text{obj}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{T}, \boldsymbol{\lambda}) \coloneqq \mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\lambda}}) + \mathcal{L}_{\text{SM}}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}, \boldsymbol{T}, \boldsymbol{\lambda}) \tag{1}$$

$\mathcal{L}_{\text{ELBO}}$ and $\mathcal{L}_{\text{SM}}$ are outlined in Eq. (2). $p_{\mathcal{D}}$ is the training data distribution.

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} \coloneqq & -\mathbb{E}_{p_{\mathcal{D}}}\left[\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},y,u)}[\log p_{\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{z}) + \log p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|y,u) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},y,u]\right] \\ \mathcal{L}_{\text{SM}} \coloneqq & \mathbb{E}_{p_{\mathcal{D}}}\left[\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},y,u)}[\|\nabla_{\boldsymbol{z}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},y,u) - \nabla_{\boldsymbol{z}} \log p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|y,u)\|^2]\right] \end{aligned} \tag{2}$$

As can be seen from the joint distribution $p(\boldsymbol{x},y,u) \propto \int_{\boldsymbol{z}} p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}|y,u)\,\mathrm{d}\boldsymbol{z}$, the prior $p(\boldsymbol{z}|y,u)$ is consistent with our SCM. In fact, based on our SCM, $p(\boldsymbol{z}|y,u)$ can be further decomposed into the product of conditional probabilities, which would result in further sparsity in $\boldsymbol{\lambda}$. This property makes our learnt model differ from existed work based on different SCMs [19,20,32], as elaborated in Appendix B.4. However, the causal graph is generally unknown *a priori*, and hence we can not pre-define the sparsity of $\boldsymbol{\lambda}$. Therefore, in our learning algorithm we treat $p(\boldsymbol{z}|y,u)$ as a generic form of prior that satisfies **Assumption 1(b)**. Without a fully specified causal graph, we use this generic prior $p(\boldsymbol{z}|y,u)$ to constrain the learning of VAE to obtain the $\boldsymbol{Z}$ without knowing their causal identities. We summarize the assumptions and identifiability results in **Theorem 1**.

Theorem 1. *Assume the data is sampled from a generative model described by*

$$p_{\boldsymbol{\xi}=(\boldsymbol{\phi},\boldsymbol{T},\boldsymbol{\lambda})}(\boldsymbol{X},\boldsymbol{Z}|Y,U) = p_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{Z})p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{Z}|Y,U),\ p_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{Z}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{X} - g_{\boldsymbol{\phi}}(\boldsymbol{Z}))$$

We have the following: (1) Under certain assumptions regarding domain variability, the model parameter $\boldsymbol{\xi}$ is identifiable up to a permutation and component-wise transformation. (2) If (1) is true and $q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},y,u)$ is flexible and positive everywhere, then our framework learns the true parameters $\boldsymbol{\xi}^*$ and true latent variables $\boldsymbol{Z}^*$ up to a permutation and component-wise transformation.

[5] Arbitrary function $\boldsymbol{\lambda}$ and sufficient statistics $\boldsymbol{T}$ are modeled by neural networks with ReLU activation due to their universal approximation ability.

We provide the detailed assumptions for (1) to hold and proof in Appendix B.2[6]. These detailed assumptions are common in identifiable VAE methods and can impact the performance of the CMB discovery in Phase II. The component-wise identifiability of $\boldsymbol{Z}$ ensures that the CMB variables and spurious variables are separable, which is a crucial prerequisite for performing CMB discovery in Phase II, as verified by empirical results in Appendix F.1. **While these assumptions may be violated in real-world applications, strict adherence to them is not always necessary.** This motivates our investigation into identifiability performance on the CMNIST dataset in Section 5.1. Empirical results demonstrate that the VAE approach we employ can achieve a decent level of identifiability, even when certain assumptions do not hold.

Phase II: Identify CMB Set. In this phase, we aim to identify the CMB representations relevant to Y among the latent variables $\boldsymbol{Z}$ we learned from the VAE framework. Traditional approaches to MB discovery involve revealing the local causal structure to the target through a series of (conditional) independence (CI) tests. However, conducting MB discovery in latent space encounters two major challenges: 1) the large dimension of the latent variable space, especially with complex and challenging input data, and 2) the mixture of variable types. These challenges can compromise the efficiency and accuracy of CMB discovery results. To tackle these two issues, we introduce an efficient search strategy and a more practical and general approach to conducting mutual-information-based (MI) independence tests. It is worth noting that we obtain CMB variables collectively by detecting and removing spurious variables with designed independence tests. We do not distinguish between the parent, child, and spouse variables within the CMB set.

Search Strategy: We find that learning the exact local graph of the target variable Y is unnecessary for identifying the CMB set. Instead of distinguishing among the CMB set, our objective is to separate spurious variables $\boldsymbol{Z}_o$ from $\boldsymbol{Z}_{cmb}$. In particularly, we discover that spurious variables $\boldsymbol{Z}_o$ and CMB variables $\boldsymbol{Z}_{cmb}$ possess different relations concerning Y conditioned on the remaining variables. If $Z_i \in \boldsymbol{Z}_o$, we have $Z_i \perp\!\!\!\perp Y | \{\boldsymbol{Z}_{\setminus i}, U\}$, where $\boldsymbol{Z}_{\setminus i} = \{\boldsymbol{Z} \setminus Z_i\}$. If $Z_i \in \boldsymbol{Z}_{cmb}$, we have $Z_i \not\perp\!\!\!\perp Y | \{\boldsymbol{Z}_{\setminus i}, U\}$. We employ mutual information as measurement for independence tests. Theoretically, we can detect spurious variables and separate them from CMB set by quantifying $\mathcal{I}(Z_i; Y | \boldsymbol{Z}_{\setminus i}, U), \forall Z_i \in \boldsymbol{Z}$. Actually, for real-world scenarios where U is unknown, it is challenging to estimate $\mathcal{I}(Z_i; Y | \boldsymbol{Z}_{\setminus i}, U)$. However, subject to our SCM structural assumptions, we have $\mathcal{I}(Z_i; Y | \boldsymbol{Z}_{\setminus i}) = \mathcal{I}(Z_i; Y | \boldsymbol{Z}_{\setminus i}, U) = 0$, if $Z_i \in \boldsymbol{Z}_o$. We provide a proof in Appendix C.1. Meanwhile, if $Z_i \in \boldsymbol{Z}_{cmb}, \mathcal{I}(Z_i; Y | \boldsymbol{Z}_{\setminus i}) > 0$. Therefore, we can separate $\boldsymbol{Z}_o$ and $\boldsymbol{Z}_{cmb}$ by performing $\mathcal{I}(Z_i; Y | \boldsymbol{Z}_{\setminus i}), \forall Z_i \in \boldsymbol{Z}$. We summarize such theoretical findings in **Proposition 1**.

[6] Our proof follows a similar approach to that of [32]. We aim to demonstrate that our assumptions on SCM do not violate the assumptions in the proof, thereby ensuring that the identifiable results hold under our framework.

Proposition 1. *If the causal graph in Figure 1 and Assumption 1 holds, we have that: 1) for* $Z_i \in \boldsymbol{Z}_o$, $Z_i \perp\!\!\!\perp Y|\boldsymbol{Z}_{\setminus i}$, $\mathcal{I}(Z_i; Y|\boldsymbol{Z}_{\setminus i}) = 0$*; 2) for* $Z_i \in \boldsymbol{Z}_{cmb}$, $Z_i \not\perp\!\!\!\perp Y|\boldsymbol{Z}_{\setminus i}$, $\mathcal{I}(Z_i; Y|\boldsymbol{Z}_{\setminus i}) > 0$.

Algorithm 1 Causal Markov Blanket Set Search Strategy

Input: M observations of $\boldsymbol{Z}$ and Y, i.e., $\{z_1(j), z_2(j), \cdots, z_N(j), y(j)\}_{j=1}^{M}$
Output: CMB set $\boldsymbol{Z}_{cmb}$
Step 1(**Fast-Search**): find the identity for each variable $Z_i \in \boldsymbol{Z}$
Initial two sets $\boldsymbol{Z}_{cmb} \leftarrow \emptyset, \boldsymbol{Z}_o \leftarrow \emptyset$
for $Z_i \in \boldsymbol{Z}, i = 1, 2, \cdots, N$, **do**
  Perform `Indtest`$(Z_i; Y|\boldsymbol{Z}_{\setminus i})$
  if $Z_i \not\perp\!\!\!\perp Y|\boldsymbol{Z}_{\setminus i}$ **then**
    $\boldsymbol{Z}_{cmb} \leftarrow \boldsymbol{Z}_{cmb} \cup Z_i$
  else
    $\boldsymbol{Z}_o \leftarrow \boldsymbol{Z}_o \cup Z_i$
  end if
end for
Step 2 (Forward-Backward Search): verify the obtained CMB set.
repeat
  for $Z_l \in \boldsymbol{Z}_{cmb}$ **do** ▷ Backward procedure
    Perform `Indtest`$(Z_l; Y|\{\boldsymbol{Z}_{cmb} \setminus Z_l\})$
    if $Z_l \perp\!\!\!\perp Y|\{\boldsymbol{Z}_{cmb} \setminus Z_l\}$ **then**
      $\boldsymbol{Z}_{cmb} \leftarrow \boldsymbol{Z}_{cmb} \setminus Z_l$
    end if
  end for
  for $Z_j \in \boldsymbol{Z}_o$ **do** ▷ Forward procedure
    Perform `Indtest`$(Z_j; Y|\boldsymbol{Z}_{cmb})$
    if $Z_j \not\perp\!\!\!\perp Y|\boldsymbol{Z}_{cmb}$ **then**
      $\boldsymbol{Z}_{cmb} \leftarrow \boldsymbol{Z}_{cmb} \cup Z_j$
    end if
  end for
until Converge

We provide the detailed proof of **Proposition 1** in Appendix C.2.

We hence propose a fast-search strategy, which performs CI tests $\mathcal{I}(Z_i; Y|\boldsymbol{Z}_{\setminus i})$ for each variable $Z_i \in \boldsymbol{Z}$. After performing the fast-search and obtaining the initial set of CMB, we verify the initial CMB set by repeatedly checking whether $\forall Z_j \in \boldsymbol{Z}_o$, $Z_j \perp\!\!\!\perp Y|\boldsymbol{Z}_{cmb}$ and $\forall Z_k \in \boldsymbol{Z}_{cmb}$, $Z_k \not\perp\!\!\!\perp Y|\{\boldsymbol{Z}_{cmb} \setminus Z_k\}$. We outline our CMB search strategy in Algorithm 1. As illustrated by **Proposition 2**, our algorithm can correctly identify the CMB variables if the assumptions hold. We provide the detailed proof for **Proposition 2** in Appendix C.3.

Proposition 2. *Assume independence test results are correct and* ***Assumption 1*** *for the proposed SCM holds, the ground truth Causal Markov Blanket set can be identified with* ***Algorithm 1****.*

In Step 1, we leverage the structural properties of CMB features to efficiently distinguish them from spurious features. However, since we condition on $N-1$ variables, which is a larger set than CMB, in step 1, we may have false positive detections and wrongly include spurious features. Thus, we employ Step 2, a forward-backward procedure, to eliminate possible errors in Step 1.

Complexity Analysis of CMB Search: Assuming it takes k iterations, with $k << N$ as observed in practice, for the verification to converge, our CMB strategy requires performing $(k+1)N$ CI tests and has a linear complexity of $\mathcal{O}(N)$. Compared to traditional MB discovery methods like IAMB [48], which directly execute the forward-backward search, our fast-search step reduces the search space and improves learning efficiency, often achieving convergence in just $k=1$ or 2 iterations in practice.

CI Tests: We use the MI-based, denoted as $\mathcal{I}$, independence tests, with a significance level α. If $\mathcal{I} < \alpha$, we declare independence. In Phase II, we need to compute the following mutual information $\mathcal{I}(Y; Z_i|\mathcal{C})$, where $\mathcal{C}$ is the set of variables to condition on. We propose a practical approach that quantifies the MI using a trained predictor and is applicable for both classification tasks and regression tasks. Calculating mutual information for classification tasks is particularly challenging due to the mixture of variable types (where Y is discrete and variables in $\boldsymbol{Z}$ are continuous). We illustrate our approach to estimating the MI in Eq. (3) and Eq. (4). By definition of mutual information and conditional entropy, we have

$$\begin{aligned}\mathcal{I}(y; z_i|\mathcal{C}) =& \mathcal{H}[y|\mathcal{C}] - \mathcal{H}[y|\{\mathcal{C}, z_i\}] \\ =& \mathbb{E}_{p(\mathcal{C})}\Big[\mathcal{H}[p(y|\mathcal{C})]\Big] - \mathbb{E}_{p(\mathcal{C},z_i)}\Big[\mathcal{H}[p(y|\mathcal{C}, z_i)]\Big] \\ =& \mathbb{E}_{p(\mathcal{C})}\Big[\mathcal{H}[\mathbb{E}_{p(\boldsymbol{z}\backslash\mathcal{C}|\mathcal{C})}[p(y|\boldsymbol{z})]]\Big] - \mathbb{E}_{p(\mathcal{C},z_i)}\Big[\mathcal{H}[\mathbb{E}_{p(\boldsymbol{z}\backslash\{\mathcal{C},z_i\}|\{\mathcal{C},z_i\})}[p(y|\boldsymbol{z})]]\Big]\end{aligned} \tag{3}$$

According to Eq. (3), we can calculate $\mathcal{I}(y; z_i|\mathcal{C})$ using a predictor distribution $p(Y|\boldsymbol{Z})$ and samples of $\boldsymbol{Z}$. For one input and its label from the training data, i.e., $\boldsymbol{x}(j), y(j), u(j) \sim p_{\mathcal{D}}$, we use the mean of learned encoder $q_{\hat{\boldsymbol{\theta}}}(\boldsymbol{Z}|\boldsymbol{X} = \boldsymbol{x}(j), Y = y(j), U = u(j))$ as the sample for $\boldsymbol{Z}$, denoted as $\boldsymbol{z}(j) = [z_1(j), z_2(j), \cdots, z_N(j)]$. After obtaining a total of M observations for $\boldsymbol{Z}$, i.e., $\{\boldsymbol{z}(j)\}_{j=1}^{M}$, we first learn a predictor $p(Y|\boldsymbol{Z})$ with $\{\boldsymbol{z}(j), y(j)\}_{j=1}^{M}$. Substituting the samples and learned distribution into Eq. (3), we have

$$\begin{aligned}\mathcal{I}(y, z_i|\mathcal{C}) \coloneqq& -\frac{1}{M}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\Big(\frac{\sum_{k=1}^{M} p\big(y=l|\mathcal{C}(j), \{\boldsymbol{z}\backslash\mathcal{C}\}(k)\big)}{M}\Big)\log\Big(\frac{\sum_{k=1}^{M} p\big(y=l|\mathcal{C}(j), \{\boldsymbol{z}\backslash\mathcal{C}\}(k)\big)}{M}\Big) + \\ & \frac{1}{M}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\Big(\frac{\sum_{k=1}^{M} p\big(y=l|\{\mathcal{C}, z_i\}(j), \{\boldsymbol{z}\backslash(\mathcal{C}, z_i)\}(k)\big)}{M}\Big)\log\Big(\frac{\sum_{k=1}^{M} p\big(y=l|\{\mathcal{C}, z_i\}(j), \{\boldsymbol{z}\backslash(\mathcal{C}, z_i)\}(k)\big)}{M}\Big)\end{aligned} \tag{4}$$

where L is the number of classes for Y. l denotes the value of Y and $l \in \{0, 1, \cdots, L-1\}$. Especially, for the fast-search strategy, $\mathcal{C} = \boldsymbol{z}_{\backslash i}$, we have:

$$\begin{aligned}\mathcal{I}(y; z_i|\boldsymbol{z}_{\backslash i}) =& -\frac{1}{M}\sum_{j=1}^{M}\sum_{l=0}^{L-1}\frac{\sum_{k=1}^{M} p\big(y=l|\boldsymbol{z}_{\backslash i}(j), z_i(k)\big)}{M}\log\frac{\sum_{k=1}^{M} p\big(y=l|\boldsymbol{z}_{\backslash i}(j), z_i(k)\big)}{M} \\ & + \frac{1}{M}\sum_{j=1}^{M}\sum_{l=0}^{L-1} p\big(y=l|\boldsymbol{z}(j)\big)\log p\big(y=l|\boldsymbol{z}(j)\big)\end{aligned} \tag{5}$$

The detailed derivation can be found in Appendix C.4. Our CMB discovery method assumes the accuracy of the `IndTest`. In practice, our approach, like other statistical tests, faces challenges with accuracy when the condition set is large and data is insufficient. Future studies aimed at improving the accuracy of independence tests could benefit from utilizing Bayesian mutual information [9].

Phase III: Construct Invariant Predictor. We utilize the obtained CMB variables $\boldsymbol{Z}_{cmb}$ and train an invariant predictor $p(Y|\boldsymbol{Z}_{cmb})$. Our training procedure is outlined in Eq. (6).

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}} \frac{1}{M} \sum_{j=1}^{M} \mathcal{L}_{\text{pred}}\Big(y(j), p_{\boldsymbol{\psi}}\big(\boldsymbol{Y}|\boldsymbol{Z}_{cmb} = \boldsymbol{z}_{cmb}(j)\big)\Big) \tag{6}$$

where $\boldsymbol{\psi}$ are the parameters of the predictor. $\mathcal{L}_{\text{pred}}(\cdot)$ represents the cross-entropy loss for classification tasks and the mean squared loss for regression tasks.

4.2 Inference Procedure

In the inference procedure, we predict labels for data from an unseen test domain, $p_{\mathcal{D}^{test}}$. One can infer from Figure 1 that $p(\boldsymbol{X}|\boldsymbol{Z})$ is domain-invariant since $X \perp\!\!\!\perp U|\boldsymbol{Z}$. Therefore, inferring the latent variable values for inputs from an unseen test domain using $p(\boldsymbol{X}|\boldsymbol{Z})$ is reasonable. We adopt the standard inference procedure from [32, 45]. Given an input from the test domain, $\boldsymbol{x}^t \sim p_{\mathcal{D}^{test}}$, we obtain the optimal values of $\boldsymbol{Z}_{cmb}$, denoted as $\boldsymbol{z}^*_{cmb}$, by solving the optimization problem outlined in Eq. (7). Here, λ_1 and λ_2 are hyperparameters used to control the scales of the learned $\boldsymbol{Z}_{cmb}$ and $\boldsymbol{Z}_o$.

$$\boldsymbol{z}^*_{cmb}, \boldsymbol{z}^*_o = \arg\min_{\boldsymbol{z}_{cmb}, \boldsymbol{z}_o} p_{\hat{\boldsymbol{\phi}}}(\boldsymbol{x}^t|\boldsymbol{z}_{cmb}, \boldsymbol{z}_o) + \lambda_1\|\boldsymbol{z}_{cmb}\|_2^2 + \lambda_2\|\boldsymbol{z}_o\|_2^2 \tag{7}$$

Then we can infer the label using the constructed invariant predictor, i.e.,

$$\hat{y} = \arg\max_{y} p_{\hat{\boldsymbol{\psi}}}(y|\boldsymbol{z}^*_{cmb}) \tag{8}$$

We provide the complexity analysis of the optimization in the inference procedure in Appendix D and the ablation study of runtime comparison in Appendix F.3.

5 Experiments

We validate our CMBRL method for OOD prediction using **synthetic and real** benchmark distribution shift datasets against state-of-the-art domain generalization baselines. We experiment on four datasets: CMNIST, CelebA, PACS, and VLCS. Our experiments, conducted across 5 trials, are summarized in tables, reporting the mean and standard deviation of accuracy. The details of datasets and implementations can be found in Appendix E.1 and E.2. Additionally, detailed ablation studies on the selection of hyperparameters are available in Appendix F.

5.1 Identifiability of Latent Variables

First, we verify that Phase I of our proposed CMBRL yields identifiable $\boldsymbol{Z}$. We compare it with VAE [30], which lacks identifiability guarantees, and iVAE [19], which uses a conditionally factorized prior. Following standard procedures [19, 20, 22, 32], we evaluate identifiability on CMNIST data by computing the average mean correlation coefficient (MCC) between latent variables recovered by different models with various random initializations. Higher MCC scores indicate stronger identifiability, as shown in Figure 3[7]. Using the same prior and training process as [32], we refer to our method as NF-iVAE. Figure 3 demonstrates that NF-iVAE recovers latent variables with better identifiability. This identifiability impacts the disentanglement of spurious and CMB variables, influencing the accuracy of the estimated CMB set. Our ablation study in Appendix F.1 empirically shows that better identifiability leads to a more accurate CMB set.

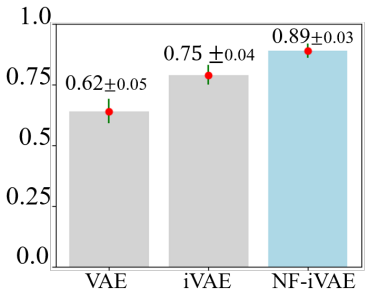

Fig. 3: MCC on CMNIST

5.2 Out-of-distribution Prediction Accuracy

We present both the in-distribution and OOD prediction accuracies of the CMNIST dataset in Table 1. First, we compare with the ERM trained on the entire training domain. Then, we compare with the popular causal invariant feature learning method, IRM, and its two variants: F-IRM (where Φ is fixed to the identity) and V-IRM (where Φ is variable). We also include the Robust MIN MAX method, which minimizes the maximum loss across multiple domains. Additionally, we compare with other approaches that leverage SCM knowledge to identify causal features of the target, including Causalrep [53], CTrans [34], and iCaRL [32]. CTrans, in particular, considers latent confounding effects between spurious representation and target and employs causal inference via intervention.

Table 1: Results on Colored MNIST in terms of accuracy (%).

Methods	Prediction Acc (%)	
	In-distribution	OOD
ERM	**85.7**±0.5	10.3±0.2
Robust MIN MAX	84.3±0.4	10.9±0.5
F-IRM GAME	63.4±1.1	60.0±2.7
V-IRM GAME	64.0±1.0	49.2±3.4
IRM	59.3±4.4	62.8±9.6
Causalrep	70.1±1.5	68.6±5.5
CTrans	76.9±0.8	72.5±1.1
iCaRL[8]	70.6±0.8	68.8±1.5
CMBRL(ours)	76.4±0.8	**74.1**±1.5

As shown in Table 1, our CMBRL method achieves optimal OOD performance while minimally compromising in-distribution accuracy. In contrast, the ERM method, which minimizes losses over the entire training domain, tends to

[7] We verify our NF-iVAE implementation's correctness by demonstrating similar identifiability performance to iCaRL in Figure 3

capture spurious statistical correlations between labels and colors. Although it achieves the highest in-distribution accuracy, there is a significant drop in prediction accuracy on OOD data. Among the multiple domain-invariant causal representation learning methods, including IRM, F-IRM GAME, and V-IRM GAME, there are notable improvements. However, the limited number of available domains (only 2) may constrain the performance of the IRM methods. Conversely, causal representation learning methods operating under various Structural Causal Models (SCMs), including our CMBRL, demonstrate superior empirical performance.

We further investigate the distinctions between the feature sets selected by iCaRL and our CMBRL. Following the approach described in [32], we first learn the causal graph between the estimated latent variables $\boldsymbol{Z}$ from NF-iVAE and Y using the PC algorithm. We then select the parent set with additional independent tests. The selected parent set yields an OOD prediction accuracy of 69.8%, consistent with the reported results in Table 1. By applying a suitable significance level α, we can select a CMB set that includes all the variables present in the selected parent set. The OOD prediction accuracy using such a CMB set increased to 74.1%. This finding supports our assumptions on the proposed SCM and suggests the presence of variables not categorized as parent variables of the target, which can further enhance OOD prediction performance. Moreover, CTrans [34] and [22] assume there is an invariant latent confounder between Y and $\boldsymbol{Z}_o$, leading to spurious correlations between Y and $\boldsymbol{X}$. For example, in an image with a water bird and sea background, the confounder could include temperature and altitude, influencing the generation of the sea (spurious variable $\boldsymbol{Z}_o$) and bird (Y). Our proposed SCM does not explicitly model latent confounders. Therefore, we cannot guarantee the generalization of our CMBRL method to scenarios involving unobserved confounders between input X and target Y. While addressing the latent confounding issue falls outside the scope of this paper, it presents an intriguing avenue for future research. We summarize empirical results on average OOD prediction accuracy over 7 settings for the CelebA dataset in Table 2. Detailed empirical results for each pair of settings can be found in Appendix H. We compare the results of our CMBRL with the ERM method and state-of-the-art causal and non-causal representation learning methods. As illustrated in Table 2, our method demonstrates optimal performance, surpassing the second-best baseline by an average margin of 4.7% across all settings. This outcome indicates the efficacy of our CMBRL methods in excluding spurious variables and avoiding reliance on spurious correlations for prediction in computer vision applications.

Table 2: Empirical results on CelebA in terms of the OOD prediction accuracy (%).

Methods	GroupDRO	MLDG	CORAL	MMD	DACNN	Mixup	ERM	CTrans	Causalrep	iCaRL	CMBRL(ours)
Avg Acc	65.1	65.6	66.7	66.2	68.7	69.5	60.7	66.3	60.4	71.5	**76.2**

Table 3: Empirical results on VLCS and PACS datasets in terms of OOD prediction accuracy (%) with a backbone of ResNet50.

Algorithms	VLCS					PACS				
	C	**L**	**S**	**V**	**Avg**	**A**	**C**	**P**	**S**	**Avg**
ERM	98.0±0.4	62.6±0.9	70.8 ±1.9	77.5 ±1.9	77.2	84.8±1.3	76.4±1.1	96.7±0.6	76.1±1.0	83.5
GroupDRO	98.1± 0.3	66.4 ± 0.9	71.0 ± 0.3	76.1 ± 1.4	77.9	83.5±0.9	79.1±0.6	96.7±0.3	78.3±2.0	84.4
MLDG	98.5± 0.3	61.7± 1.2	73.6± 1.8	75.0± 0.8	77.2	85.5±1.4	80.1±1.7	97.4±0.3	76.6±1.1	84.9
CORAL	96.9± 0.9	65.7± 1.2	73.3± 0.7	78.7± 0.8	78.7	88.3±0.2	80.0±0.7	97.5±0.3	78.8±1.3	86.2
MMD	98.3± 0.1	65.6± 0.7	69.7± 1.0	75.7± 0.9	77.3	86.1±1.4	79.4±0.9	96.6±0.2	76.5±0.7	84.6
RSC	97.5± 0.6	63.1± 1.2	73.0± 1.3	76.2± 0.5	77.5	85.4±0.8	79.7±1.8	97.6±0.3	78.2±1.2	85.2
Mixup	98.4± 0.3	63.4± 0.7	72.9± 0.8	76.1± 1.2	77.7	86.1±0.7	78.9±0.8	97.6±0.1	75.8±1.8	84.6
DANN	98.5± 1.3	64.9± 1.3	72.6± 1.4	78.7± 1.7	78.2	86.4±0.8	77.4±0.8	97.3±0.4	73.5±2.3	83.6
CDANN	97.6± 0.6	65.2± 0.8	73.4± 1.4	76.9± 0.5	78.3	84.6±1.8	75.5±0.9	96.8±0.3	73.5±0.6	82.6
MTL	97.6± 0.6	60.6± 1.3	71.0± 1.2	77.2± 0.7	76.6	87.5±0.8	77.1±0.7	96.4±0.8	77.3±1.8	84.6
ARM	97.2± 0.5	62.7± 1.	70.6± 0.6	75.8± 0.9	76.6	86.8±0.6	76.8±0.7	97.4±0.3	79.3±1.2	85.1
IRM	**98.6**±0.1	66.0 ±0.9	72.3 ±0.6	77.3 ±0.9	78.5	84.7±0.4	80.0±0.6	97.2±0.3	79.3±1.0	85.5
SagNet	97.3± 0.4	61.6± 0.8	73.4± 1.9	77.6± 0.4	77.5	87.4±1.0	80.7±0.6	97.1±0.1	80.0±0.4	86.3
iCaRL	-	-	-	-	81.8	-	-	-	-	88.7
CMBRL (ours)	**98.7**±0.2	**71.2**±0.1	**77.1**±0.4	**82.1**±0.1	**82.3**	**89.2**±0.7	**85.3**±1.2	**97.7**±0.5	**84.1**±0.6	**89.1**

We present the empirical results for PACS and VLCS regarding OOD prediction accuracy in Table 3. These datasets serve as benchmarks for domain generalization tasks. For a comprehensive comparison, we evaluate not only against causal-inspired methods but also against various other DG methods, including GroupDRO [41], MLDG [25], CORAL [44], RSC [15], Mixup [54], SageNet [36], and others. We adhere to the experimental settings outlined in [13]. We report the OOD performance for each pair of train and test domains and the average accuracy across all combinations. As depicted in Table 3, our CMBRL method establishes state-of-the-art performance compared to popular alternatives in the domain generalization landscape.

6 Conclusion

In this work, we investigate the Causal Markov Blanket discovery for high-dimensional data. We establish a framework guided by an SCM describing the data generation process, allowing for the Causal Markov Blanket discovery in latent space. We then construct an invariant prediction mechanism utilizing the CMB representations, which is suitable for OOD prediction. This framework can be further employed for causal analysis, reasoning, and intervention. We propose a three-phase algorithm to disentangle CMB features from spurious ones, reducing the risk of the prediction model relying on spurious correlations. It is worth noting that our method requires the satisfaction of assumptions on SCM and the assurance of the identifiability of latent variables. It may also suffer from inefficiencies due to the three training phases and optimization during inference. However, our method demonstrates its effectiveness by achieving significant OOD prediction performance, surpassing state-of-the-art causal representation learning methods and domain generalization methods on multiple benchmark distribution shift datasets.

Acknowledgements

This work was supported in part by the National Science Foundation award IIS 2236026 and in part by IBM through the IBM-Rensselaer Future of Computing Research Collaboration.

References

1. Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., Rish, I.: Invariance principle meets information bottleneck for out-of-distribution generalization. In: Advances in Neural Inf. Proc. Systems (2021) 3
2. Ahuja, K., Shanmugam, K., Varshney, K., Dhurandhar, A.: Invariant risk minimization game. In: International Conference on Machine Learning (2020) 3
3. Ahuja, K., Wang, J., Dhurandhar, A., Shanmugam, K., Varshney, K.R.: Empirical or invariant risk minimization? a sample complexity perspective. In: International Conference on Learning Representations (2021) 3, 5
4. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. Journal of Machine Learning Research **11**(1) (2010) 2, 3
5. Aliferis, C.F., Tsamardinos, I., Statnikov, A.: Hiton: a novel markov blanket algorithm for optimal variable selection. In: AMIA annual symposium proceedings. vol. 2003, p. 21. American Medical Informatics Association (2003) 3
6. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019) 3
7. Cui, P., Athey, S.: Stable learning establishes some common ground between causal inference and machine learning. Nature Machine Intelligence **4**(2), 110–115 (2022) 3
8. Cui, P., Shen, Z., Li, S., Yao, L., Li, Y., Chu, Z., Gao, J.: Causal inference meets machine learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3527–3528 (2020) 3
9. Cui, Z., Yin, N., Wang, Y., Ji, Q.: Empirical bayesian approaches for robust constraint-based causal discovery under insufficient data. International Joint Conference on Artificial Intelligence (2022) 11
10. Fu, S., Desmarais, M.C.: Fast markov blanket discovery algorithm via local learning within single pass. In: Advances in Artificial Intelligence: 21st Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2008 Windsor, Canada, May 28-30, 2008 Proceedings 21. pp. 96–107. Springer (2008) 3
11. Gao, T., Ji, Q.: Local causal discovery of direct causes and effects. Advances in Neural Information Processing Systems **28** (2015) 3
12. Gao, T., Wang, Z., Ji, Q.: Structured feature selection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015) 2, 3, 4
13. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv preprint arXiv:2007.01434 (2020) 14
14. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019) 3
15. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: European Conference on Computer Vision. pp. 124–140. Springer (2020) 14

16. Hyvärinen, A., Dayan, P.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**(4) (2005) 7
17. Janzing, D.: Causal regularization. Advances in Neural Information Processing Systems **32** (2019) 3
18. Jiang, Y., Veitch, V.: Invariant and transportable representations for anti-causal domain shifts. arXiv preprint arXiv:2207.01603 (2022) 3
19. Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: International Conference on Artificial Intelligence and Statistics. pp. 2207–2217. PMLR (2020) 3, 4, 7, 12
20. Khemakhem, I., Monti, R., Kingma, D., Hyvarinen, A.: Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. Advances in Neural Information Processing Systems **33**, 12768–12778 (2020) 4, 7, 12
21. Koller, D., Sahami, M.: Toward optimal feature selection. Tech. rep., Stanford InfoLab (1996) 3
22. Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., Zhang, K.: Partial identifiability for domain adaptation. arXiv preprint arXiv:2306.06510 (2023) 2, 3, 5, 12, 13
23. Koyama, M., Yamaguchi, S.: When is invariance useful in an out-of-distribution generalization problem? arXiv preprint arXiv:2008.01883 (2020) 3
24. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018) 3
25. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5542–5550 (2017) 14
26. Li, Z., Cai, R., Chen, G., Sun, B., Hao, Z., Zhang, K.: Subspace identification for multi-source domain adaptation. Advances in Neural Information Processing Systems **36** (2024) 3
27. Liu, B., Wang, D., Yang, X., Zhou, Y., Yao, R., Shao, Z., Zhao, J.: Show, deconfound and tell: Image captioning with causal inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18041–18050 (2022) 4
28. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019) 3
29. Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., Tschannen, M.: Weakly-supervised disentanglement without compromises. In: International Conference on Machine Learning. pp. 6348–6359. PMLR (2020) 3
30. Lopez, R., Regier, J., Jordan, M.I., Yosef, N.: Information constraints on auto-encoding variational bayes. Advances in neural information processing systems **31** (2018) 12
31. Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., Bottou, L.: Discovering causal signals in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6979–6987 (2017) 2
32. Lu, C., Wu, Y., Hernández-Lobato, J.M., Schölkopf, B.: Invariant causal representation learning for out-of-distribution generalization. In: International Conference on Learning Representations (2021) 2, 3, 4, 5, 7, 8, 11, 12, 13
33. Mao, C., Cha, A., Gupta, A., Wang, H., Yang, J., Vondrick, C.: Generative interventions for causal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3947–3956 (2021) 2, 3, 4, 5

34. Mao, C., Xia, K., Wang, J., Wang, H., Yang, J., Bareinboim, E., Vondrick, C.: Causal transportability for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7521–7531 (2022) 2, 4, 5, 12, 13
35. Nagarajan, V., Andreassen, A., Neyshabur, B.: Understanding the failure modes of out-of-distribution generalization. arXiv preprint arXiv:2010.15775 (2020) 1
36. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8690–8699 (2021) 14
37. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann (1988) 4
38. Pellet, J.P., Elisseeff, A.: Using markov blankets for causal structure learning. Journal of Machine Learning Research **9**(7) (2008) 2, 3
39. Peters, J., Buhlmann, P., Meinshausen, N.: Causal inference using invariant prediction: identification and confidence intervals. arxiv. Methodology (2015) 2, 5
40. Rosenfeld, E., Ravikumar, P., Risteski, A.: The risks of invariant risk minimization. In: International Conference on Learning Representations (2021) 3
41. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019) 2, 14
42. Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., Zhang, T.: Weakly supervised disentangled generative causal representation learning. The Journal of Machine Learning Research **23**(1), 10994–11048 (2022) 3
43. Shu, Y., Cao, Z., Wang, C., Wang, J., Long, M.: Open domain generalization with domain-augmented meta-learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9624–9633 (2021) 3
44. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 443–450. Springer (2016) 14
45. Sun, X., Wu, B., Zheng, X., Liu, C., Chen, W., Qin, T., Liu, T.y.: Latent causal invariant model. arXiv preprint arXiv:2011.02203 (2020) 11
46. Talon, D., Lippe, P., James, S., Del Bue, A., Magliacane, S.: Towards the reusability and compositionality of causal representations. arXiv preprint arXiv:2403.09830 (2024) 3
47. Tan, Y., Liu, Z.: Feature selection and prediction with a markov blanket structure learning algorithm. In: BMC bioinformatics. vol. 14, pp. 1–3. BioMed Central (2013) 2, 3
48. Tsamardinos, I., Aliferis, C.F., Statnikov, A.R., Statnikov, E.: Algorithms for large scale markov blanket discovery. In: FLAIRS conference. vol. 2, pp. 376–380. St. Augustine, FL (2003) 3, 10
49. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation **23**(7), 1661–1674 (2011) 7
50. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. Advances in neural information processing systems **31** (2018) 3
51. Wang, H., Xiao, C., Kossaifi, J., Yu, Z., Anandkumar, A., Wang, Z.: Augmax: Adversarial composition of random augmentations for robust training. Advances in neural information processing systems **34**, 237–250 (2021) 3

52. Wang, T., Huang, J., Zhang, H., Sun, Q.: Visual commonsense r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10760–10770 (2020) 2, 4
53. Wang, Y., Jordan, M.I.: Desiderata for representation learning: A causal perspective. arXiv preprint arXiv:2109.03795 (2021) 12
54. Yan, S., Song, H., Li, N., Zou, L., Ren, L.: Improve unsupervised domain adaptation with mixup training. arXiv preprint arXiv:2001.00677 (2020) 14
55. Yin, N., Wang, H., Gao, T., Dhurandhar, A., Ji, Q.: Causal markov blanket representation learning for out-of-distribution generalization. In: Causal Representation Learning Workshop at NeurIPS 2023 (2023) 2, 4
56. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019) 3
57. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 3