Causal Markov Blanket Representation Learning for Out-of-distribution Generalization

Naiyu Yin
Rensselaer Polytechnic Institute
yinn2@rpi.edu

Hanjing Wang
Rensselaer Polytechnic Institute
wangh36@rpi.edu

Tian Gao
IBM Research
tgao@us.ibm.com

Amit Dhurandhar
IBM Research
adhuran@us.ibm.com

Qiang Ji
Rensselaer Polytechnic Institute
jiq@rpi.edu

Abstract

The pursuit of generalizable representations in the realm of machine learning and computer vision is a dynamic field of research. Typically, current methods aim to secure invariant representations by either harnessing domain expertise or leveraging data from multiple domains. In this paper, we introduce a novel approach that involves acquiring Causal Markov Blanket (CMB) representations to improve prediction performance in the face of distribution shifts. Causal Markov Blanket representations comprise the direct causes and effects of the target variable. Theoretical analyses have demonstrated their capacity to harbor maximum information about the target, resulting in minimal Bayes error during prediction. To elaborate, our approach commences with the introduction of a novel structural causal model (SCM) equipped with latent representations, designed to capture the underlying causal mechanisms governing the data generation process. Subsequently, we propose a CMB representation learning framework that derives representations conforming to the proposed SCM. In comparison to state-of-the-art domain generalization methods, our approach exhibits robustness and adaptability under distribution shifts.

1 Introduction

Despite the remarkable advancements achieved by deep learning models in various real-world applications, they are still plagued by certain limitations, including their susceptibility to poor out-of-distribution (OOD) performance, a lack of interpretability, and fairness concerns. In particular, regarding OOD generalization, both theoretical analysis and empirical evidence have established that the primary cause of failure stems from erroneous associations between irrelevant features and the target prediction [Nagarajan et al., 2020]. These spurious associations arise due to data biases and can vary as data distributions shift. If left unaddressed, these issues can not only result in significant predictive errors but also raise serious ethical concerns in critical tasks like autonomous driving, crime prediction, and personalized medicine.

Numerous efforts have been devoted to tackling these limitations, as evidenced by prior research [Ajakan et al., 2014, Arjovsky et al., 2019, Blanchard et al., 2021, Li et al., 2018, Carlucci et al., 2019, Mao et al., 2021, Huang et al., 2020, Qiao et al., 2020]. However, these endeavors have demonstrated varying degrees of success, often relying on additional assumptions or data from multiple domains, which can be impractical in real-world applications. To enhance out-of-distribution (OOD) prediction performance, we have adopted a novel approach grounded in causal learning, formulating the prediction task from a causal perspective. Causal learning entails the utilization of a

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

structural causal model (SCM) to capture the underlying data generation mechanism, encapsulating the intrinsic, stable, and interpretable relationships within the data. Instead of depending on superficial statistical correlations, our framework harnesses robust causal relationships that remain invariant amidst distribution shifts, thus proving highly effective in OOD scenarios. Precisely, our approach aims for the acquisition of Causal Markov Blanket (CMB) features, which consist of parents, children, and spouses of the target variable. These CMB features possess a theoretical guarantee of being invariant and interpretable when it comes to predicting the target variable. Under the purview of our proposed SCM, the influence of spurious features is mitigated by conditioning on the CMB features. This prevents the model from exploiting spurious features for prediction. Notably, in contrast to previous works [Subbaswamy et al., 2019, Peters et al., 2015, Kyono et al., 2020, Lu et al., 2021, Mao et al., 2022], which primarily focus on selecting parent or child variables, our framework is designed to acquire the CMB features that encompass completeness, high productivity, and strong invariance across domains. These optimal properties render the CMB features theoretically well-suited for prediction tasks.

The main contributions of our work are as follows: 1) We employ a novel SCM to capture intricate factors and their causal relations that underline the data generation mechanism. 2) We introduce a training framework aimed at learning the Causal Markov Blanket representations. 3) We showcase the efficacy of our proposed approach on benchmark datasets with distribution shifts. Our method demonstrates enhanced generalization performance in the presence of these shifts.

2 Related-work

Within this section, we will undertake a review of previous studies that have focused on the acquisition of representations capable of achieving generalization across various domains. Specifically, we will explore two main categories: causal approaches and non-causal approaches.

Causal approaches: Causal approaches can be categorized into two types, depending on whether interventions are carried out. Methods without intervention encompass stable representation learning approaches [Cui et al., 2020, Cui and Athey, 2022, Janzing, 2019, Jiang and Veitch, 2022] and invariant feature learning methods [Arjovsky et al., 2019, Koyama and Yamaguchi, 2020, Ahuja et al., 2020, 2021b, Rosenfeld et al., 2021, Ahuja et al., 2021a]. For stable representation learning methods, the goal is to acquire causal or anti-causal features through strategies like covariate balancing or by employing SCM as a form of regularization. Invariant feature learning methods, on the other hand, seek to learn features that remain invariant according to specific invariance criteria from multi-environmental data. One widely explored approach in this context is invariant risk minimization (IRM). Subsequent research has led to more efficient variants [Ahuja et al., 2020] and further theoretical analysis [Ahuja et al., 2021b]. However, recent findings have revealed limitations in certain cases [Rosenfeld et al., 2021, Ahuja et al., 2021a], where it fails to discover such predictors. Other strategies include risk variance regularization [Krueger et al., 2021], domain gradient alignments [Koyama and Yamaguchi, 2020], smoothing cross-domain interpolation paths [Chuang and Mroueh, 2021], and task-oriented techniques [Zhang et al., 2021]. Nevertheless, these approaches often require information that distinctly distinguishes various domains or some level of target domain information, which can be challenging to obtain in real-world applications. Causal learning methods with intervention encompass robust feature learning through data augmentation and transportable interventional inference-guided feature learning techniques. In the case of Mao et al. [2021], intervention occurs at the input level by identifying a set of transformations that can be applied without altering the invariant features. However, the selection of permissible transformations necessitates domain expertise. On the other hand, Liu et al. [2022], Wang et al. [2020], and Mao et al. [2022] estimate the invariant and transportable interventional distribution between input and target through backdoor/frontdoor adjustment. Nonetheless, these approaches require the identification and estimation of all covariation sources between input and target, limiting their applicability in real-world scenarios. Although Mao et al. [2022] avoids this issue by employing front door adjustment, it enlarges computational complexity during training due to the integration over input.

Non-Causal approaches: Non-causal methods typically involve generating new data and applying data augmentation techniques. These methods encompass disentangled representation learning [Khemakhem et al., 2020, Locatello et al., 2020, 2019, Shen et al., 2022], “mix-up" strategies [Zhang et al., 2017, Yun et al., 2019, Hendrycks et al., 2019], adversarial training techniques[Volpi et al., 2018, Wang et al., 2021], and frequency spectrum-based strategies [Sun et al., 2021, Zhang et al.,

2023]. However, it's important to note that these strategies are often heuristic in nature and can be computationally expensive, especially when adversarial training is required.

3 Causal Markov Blanket Representation Learning

This paper focuses on tasks aimed at ensuring accurate predictions in the face of distributional shifts. Our approach involves formulating, analyzing, and addressing these tasks through a causal perspective. We integrate our comprehension of the data generation process into a structural causal model (SCM) and utilize causal methodologies to generate prediction distributions that can be leveraged for out-of-distribution prediction. We demonstrate that within our SCM, the prediction distribution utilizing Causal Markov Blanket (CMB) variables remains invariant and transportable, making it well-suited for prediction under distribution shifts.

3.1 Causal Markov Blanket Features

Causal Markov Blanket (CMB) features are pivotal in pinpointing relevant variables for predicting targets with higher accuracy despite shifts in distribution. The definition of Markov Blanket is outlined in **Definition 3.1**.

Definition 3.1 (**Markov Blanket [Pearl, 1988]**). *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 $Z \perp\!\!\!\perp T | \boldsymbol{MB}_T, \forall Z \in \{\boldsymbol{V} \backslash T \backslash \boldsymbol{MB}_T\}$.*

Essentially, the $\mathbf{MB}_T$ encompasses **the parent, children, and spouse variables** of the target T. As highlighted in [Gao et al., 2015], the CMB features of the target possess two 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.

3.2 Structural Causal Model of the Data Generation Process

Causal representation learning [Peters et al., 2015, Lu et al., 2021, Mao et al., 2022] seeks to discern the foundational causal mechanisms for the data generation process in a directed acyclic graph (DAG). Advancing and enhancing the SCMs from prior works, we present a novel SCM, as illustrated in Figure 1, which integrates CMB variables. In the proposed SCM, $\boldsymbol{X}$ stands for the high-dimensional input data, such as images, videos, or texts. Y is the target variable for prediction, while $\boldsymbol{Z} = \{\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s, \boldsymbol{Z}_o\}$ indicates the latent, high-level multi-dimensional representations for generating $\boldsymbol{X}$. We categorize the latent representations $\boldsymbol{Z}$ into four types: $\boldsymbol{Z}_p$ represent the parent variables of Y; $\boldsymbol{Z}_c$ correspond to the children variables of Y; $\boldsymbol{Z}_s$ are the spouse variables of Y, and $\boldsymbol{Z}_o$ relate to the spurious variables for predicting Y. The CMB representations are denoted as $\boldsymbol{Z}_{cmb} = \{\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s\}$. To model the distribution shifts, we introduce the domain-specific latent variable, coin U_x. It denotes any information specific to the domain. We assume that there is no hidden confounder between $\boldsymbol{Z}_o$ and Y.

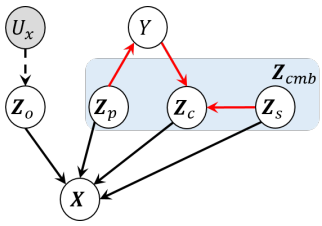

Figure 1: The proposed SCM.

Invariant prediction mechanism: According to the SCM in Figure 1, U_x interacts with $\boldsymbol{X}$ through the spurious representation $\boldsymbol{Z}_o$, i.e., $U_x \rightarrow \boldsymbol{Z}_o \rightarrow \boldsymbol{X}$, and affects Y via the pathways $U_x \rightarrow \boldsymbol{Z}_o \rightarrow \boldsymbol{X} \leftarrow \boldsymbol{Z}_c \leftarrow Y$ and $U_x \rightarrow \boldsymbol{Z}_o \rightarrow \boldsymbol{X} \leftarrow \boldsymbol{Z}_p \rightarrow Y$ when $\boldsymbol{X}$ is given. We denote the observational source distribution as π^s and the target distribution as π^t. When there's a shift in distribution, it results in $p^s(U_x) \neq p^t(U_x)$. $p(Y|\boldsymbol{X})$, estimated by the traditional classifier, unfortunately, captures the covariance between U_x, $\boldsymbol{X}$, and Y, resulting in $p^s(Y|\boldsymbol{X}) \neq p^t(Y|\boldsymbol{X})$. It motivates us to identify the CMB variables of Y, whereby giving CMB variables, the influence from the domain variable U_x is blocked, i.e., $p^s(Y|\boldsymbol{Z}_{cmb}) = p^t(Y|\boldsymbol{Z}_{cmb})$. Moreover, $\boldsymbol{Z}_{cmb}$ yields maximum information of Y compared to any of its subsets, rendering better prediction performance. Hence, this paper aims to discern the CMB representations and leverage $p(Y|\boldsymbol{Z}_{cmb})$ for generalizable prediction.

Comparision between SCMs: Compared to the distinct SCM employed by prior works [Peters et al., 2015, Lu et al., 2021, Mao et al., 2022], our SCMs share certain characteristics: 1) the input $\boldsymbol{X}$ can be generated through latent-level factors $\boldsymbol{Z}$; 2) distribution shifts arise from the domain-specific variables U_x. U_x induces changes in the distribution of input $\boldsymbol{X}$ by influencing the distribution

of domain-variant spurious variables $\boldsymbol{Z}_o$. In practical scenarios, the latent representations $\boldsymbol{Z}$ are often observable, and only their transformation $\boldsymbol{X}$ is observable. Consequently, methods that seek causal variables from the observed set of $\boldsymbol{Z}$, including Invariant Causal Prediction (ICP) [Peters et al., 2015] and many state-of-art Markov Blanket discovery methods [Pellet and Elisseeff, 2008, Aliferis et al., 2010, Tan and Liu, 2013], may falter when applied to $\boldsymbol{X}$. In particular, directly applying MB discovery methods to each pixel in $\boldsymbol{X}$ is not only computationally expansive but may not be able to disentangle CMB variables from spurious variables in scenarios that every pixel is influenced by all four types of latent variables [Lu et al., 2021]. However, existing methods mostly assume that causal features are the parent variables to target. They ignore the children and spouses variables that are also predictive of targets and result in the covariate between U_x, $\boldsymbol{X}$, and Y. However, our SCM does not account for the following scenarios: 1) latent confounder between $\boldsymbol{Z}_o$ and Y; 2) causal relations within $\boldsymbol{Z}$; 3) domain variable U_x's influence on $\boldsymbol{Z}_{cmb}$ variables. Addressing these issues may be interesting for future work but is not within the scope of this paper.

Examples on images: To shed more light on our idea, we use an image from the colored MNIST dataset as an exemplar, elucidating the notions of representations $\boldsymbol{Z}_{cmb}$ and $\boldsymbol{Z}_o$. In the realm of image classification, $\boldsymbol{Z}_{cmb}$ generally encapsulates attributes inherent to the object, like the digit's shape [Lopez-Paz et al., 2017]. Conversely, $\boldsymbol{Z}_o$ gleans details from other aspects of the image, such as the background color information. Consider an image depicting a digit 1 on a red background. In this context, $\boldsymbol{Z}_o$ represents the features responsible for generating the red background characteristics. Models that hinge on statistical dependency will inherently detect the co-occurrence of the red background $\boldsymbol{Z}_o$ and the "digit 1" label Y. Thus, faced with a distributional shift, such as predicting the digit 1 on a blue background, the model, having depended on the red background to identify the digit, might fail.

3.3 Causal Markov Blanket Representation Learning and Inference

We formulate the CMB representation learning problem into a SCM learning framework. Under the framework, we aim to learn the causal mechanism with a known causal structure. Initially, we parameterize the SCM using conditional distributions, which we subsequently model through neural networks. The parameters of these networks are determined by minimizing the logarithmic marginal likelihood of observed variables, denoted as $-\log p(\boldsymbol{X}, Y)$. Ultimately, we construct an invariant prediction mechanism $p(Y|\boldsymbol{Z}_{cmb})$, allowing us to make inferences about Y.

3.3.1 SCM Parameterization

To parameterize the proposed SCM, we decompose the joint distribution of all variables employing the Bayesian network chain rule, as depicted in Eq. (1):

$$p(\boldsymbol{x}, y, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, u_x) = p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_p)p(\boldsymbol{z}_s)p(y|z_p)p(\boldsymbol{z}_c|y, z_s)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \tag{1}$$

Notably, we employ Bayes' theorem and sum/product rules to reformulate Eq. (1) as follows:

$$p(\boldsymbol{x}, y, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, u_x) = p(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_o|u_x)p(u_x)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \tag{2}$$

The derivation of Eq. (2) is provided in Appendix A. Given that $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o$ remain unobserved, we utilize expressive neural networks to parameterize their respective conditional distributions. Specifically, we characterize $p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o)$ using a decoder parameterized by Φ, and $p(y|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)$ through a classifier parameterized by Ψ. We treat $p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s), p(\boldsymbol{z}_o|u_x), p(u_x)$ as prior distributions. Moreover, to deduce the unobserved representations $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o$, we introduce a variational distribution $q(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o|\boldsymbol{x})$ to approximate $p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o|\boldsymbol{x})$. For simplicity, we further assume that $q(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o|\boldsymbol{x}) = q(\boldsymbol{z}_p|\boldsymbol{x})q(\boldsymbol{z}_c|\boldsymbol{x})q(\boldsymbol{z}_s|\boldsymbol{x})q(\boldsymbol{z}_o|\boldsymbol{x})$ and parameterize the four encoders, $q(\boldsymbol{z}_p|\boldsymbol{x}), q(\boldsymbol{z}_c|\boldsymbol{x}), q(\boldsymbol{z}_s|\boldsymbol{x})$, and $q(\boldsymbol{z}_o|\boldsymbol{x})$, with Θ_p, Θ_c, Θ_s, and Θ_o, respectively.

3.3.2 SCM causal mechanism Learning

Given the SCM parameterization, we deduce the parameters Θ, Φ, Ψ by minimizing the training objective $\mathcal{L}_{nll}$ defined in Eq. (3), which is an upper bound of $-\log p(\boldsymbol{x}, y)$. We provide the detailed derivation of Eq. (3) in Appendix B:

$$\begin{aligned}\mathcal{L}_{nll} \coloneqq &- \mathbb{E}_{q_{\{\Theta_p,\Theta_c,\Theta_s\}}(\boldsymbol{z}_p,\boldsymbol{z}_c,\boldsymbol{z}_s|\boldsymbol{x})}[\log p_\Psi(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)] - \mathbb{E}_{q_\Theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\Phi(\boldsymbol{x}|\boldsymbol{z})] \\ &+ KL\Big(q_{\{\Theta_p,\Theta_c,\Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)\Big) + KL\Big(q_{\Theta_o}(\boldsymbol{z}_o|\boldsymbol{x})||p(\boldsymbol{z}_o)\Big)\end{aligned} \tag{3}$$

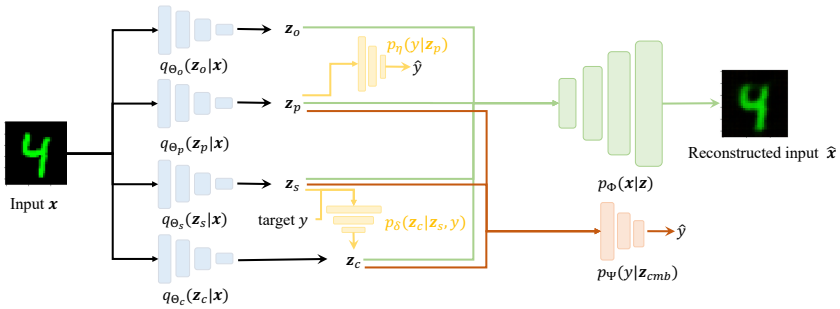

Figure 2: The proposed Causal Markov Blanket representation learning framework

where $\Theta_{cmb} = \{\Theta_p, \Theta_c, \Theta_s\}$. As depicted in Figure 2, we postulate that $q(\boldsymbol{z}_p|\boldsymbol{x}), q(\boldsymbol{z}_c|\boldsymbol{x}), q(\boldsymbol{z}_s|\boldsymbol{x})$ and $q(\boldsymbol{z}_o|\boldsymbol{x})$ all follow Gaussian distributions. The neural networks directly output the trainable mean and covariance matrices. The distributions $p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)$ and $p(\boldsymbol{z}_o)$ are the prior distributions for the CMB variables and spurious variables. With no external knowledge regarding these high-level latent variables provided, we leverage the known causal structure and posit self-defined distributions for the priors $\boldsymbol{z}_o, \boldsymbol{z}_p, \boldsymbol{z}_c$, and $p(\boldsymbol{z}_s)$. We assume that $p(\boldsymbol{z}_p)$, $p(\boldsymbol{z}_c)$ and $p(\boldsymbol{z}_s)$ follow normal distribution $\mathcal{N}(0, \boldsymbol{I})$[1]. In particular, $\boldsymbol{z}_o$ fluctuates with alterations in u_x, mirroring domain-specific impacts. This motivates us to represent $p(\boldsymbol{z}_o)$ with a mixture Gaussian distribution, i.e., $p(\boldsymbol{z}_o) = \sum_{u_x} p(\boldsymbol{z}_o|u_x)p(u_x)$. Here, we adopt a common simplification procedure to assume U_x is a discrete variable. $p(u_x)$ follows a categorical distribution with a pre-defined class count J. $p(\boldsymbol{z}_o|u_x)$ is a Gaussian distribution that exhibits unique means and covariance matrices for diverse u_x values. While $p(\boldsymbol{z}_o|u_x)$ and $p(u_x)$ aren't directly learned through neural networks, owing to the absence of supervision for $\boldsymbol{z}_o$ and u_x, they undergo iterative updates during each training epoch. Essentially, we cluster the prevailing values of $\boldsymbol{z}_o$ for the training data into J bins by the Kmeans method and subsequently refine $p(\boldsymbol{z}_o|u_x)$ based on the samples within each cluster. **CMB constraint:** While the loss function in Eq. (3) proves effective, it doesn't ensure that the obtained $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s$ are the variables subject to our proposed SCM. This limitation arises primarily because $\mathcal{L}_{nll}$ emphasizes introducing $p_\Psi(y|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)$ for direct inference, sidelining the underlying mechanics within the subgraph of $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s$. However, the acquisition of CMB features is both sufficient and necessary to achieve domain-invariant classification, as these features collectively capture the most pertinent information concerning the target variable y. Nevertheless, to bolster disentanglement within CMB set, we integrate two additional neural networks that adhere to the causal interactions among $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s$: one network models $p_\eta(y|\boldsymbol{z}_p)$, parameterized by η, and another reconstructs $\boldsymbol{z}_c$ given $\boldsymbol{z}_s$ and y, represented as $p_\delta(\boldsymbol{z}_c|\boldsymbol{z}_s, y)$. The CMB constraint $\mathcal{R}_{cmb}$ is:

$$\mathcal{R}_{cmb} = -\mathbb{E}_{q_{\Theta_p}(\boldsymbol{z}_p|\boldsymbol{x})}[\log p_\eta(y|\boldsymbol{z}_p)] - \mathbb{E}_{q_{\{\Theta_c,\Theta_s\}}(\boldsymbol{z}_c,\boldsymbol{z}_s|\boldsymbol{x})}[\log p_\delta(\boldsymbol{z}_c|\boldsymbol{z}_s, y)] \tag{4}$$

We aim to solve the following problems during training:

$$\Theta^*, \Phi^*, \Psi^*, \eta^*, \delta^* = \arg\min_{\Theta,\Phi,\Psi,\eta,\delta} \mathcal{L}_{obj}, \quad \mathcal{L}_{obj} = \mathcal{L}_{nll} + \lambda_1 \mathcal{R}_{cmb} \tag{5}$$

where λ_1 is non-negative coefficients and serves to balance the respective loss functions. We outline our detailed training procedure in Algorithm 1.

3.3.3 Inference procedure

For OOD prediction, we aim to infer the labels for data from an unseen test domain using the learned distribution $p_{\Phi^*}(\boldsymbol{X}|\boldsymbol{Z})$ and $p_{\Psi^*}(Y|\boldsymbol{Z}_{cmb})$. Since $\boldsymbol{X} \perp\!\!\!\perp U_X|\boldsymbol{Z}$, the generative mechanism $p(\boldsymbol{X}|\boldsymbol{Z})$ is invariant across domains. Following Lu et al. [2021], we first infer $\boldsymbol{z}^t_{cmb}$ for an input $\boldsymbol{x}^t$ from unseen test domain via $p_{\Phi^*}(\boldsymbol{X}|\boldsymbol{Z})$, as outlined in Eq. (6). λ_2 and λ_3 are the hyperparameters to control the scales of the learned $\boldsymbol{Z}_{cmb}, \boldsymbol{Z}_o$.

$$\boldsymbol{z}^*_{cmb}, \boldsymbol{z}^*_o = \arg\min_{\boldsymbol{z}_{cmb},\boldsymbol{z}_o} p_{\Phi^*}(\boldsymbol{x}^t|\boldsymbol{z}_{cmb}, \boldsymbol{z}_o) + \lambda_2\|\boldsymbol{z}_{cmb}\|_2^2 + \lambda_3\|\boldsymbol{z}_o\|_2^2 \tag{6}$$

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

$$\hat{y} = \arg\max_y p_{\Psi^*}(y|\boldsymbol{z}^*_{cmb}) \tag{7}$$

[1] $\boldsymbol{I}$ is the identity matrix.

Algorithm 1 Causal Markov Blanket Representation Learning Procedure

Input: Training set $\mathcal{D}$ over $\{(\boldsymbol{X}, Y)\}$; $p(U_x)$; $|U_x| = J$.
Goal: Estimate $q_{\Theta_p}(\boldsymbol{Z}_p|\boldsymbol{X})$, $q_{\Theta_c}(\boldsymbol{Z}_c|\boldsymbol{X})$, $q_{\Theta_s}(\boldsymbol{Z}_s|\boldsymbol{X})$, $q_{\Theta_o}(\boldsymbol{Z}_o|\boldsymbol{X})$, $p_\Phi(\boldsymbol{X}|\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s, \boldsymbol{Z}_o)$, $p_\Psi(Y|\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s)$, $p_\eta(Y|\boldsymbol{Z}_p)$, and $p_\delta(\boldsymbol{Z}_c|\boldsymbol{Z}_s, Y)$
Initialize encoders, decoder, and classifier.
for $i = 1, 2, \cdots, M$ **do**
    Obtain one input observation $\boldsymbol{x}^i$
    Input $\boldsymbol{X} = \boldsymbol{x}^i$ into the encoder $q_{\Theta_o}(\boldsymbol{Z}_o|\boldsymbol{X})$, obtain $\boldsymbol{z}_o^i = \boldsymbol{\mu}_{\boldsymbol{z}_o}(\boldsymbol{x}^i)$
end for
Cluster $\{\boldsymbol{z}_o^i\}_{i=1}^M$ into J bins with Kmeans algorithm. For each bin $U = u_j$, we set the mean and variance for $p(\boldsymbol{Z}_o|U_x = u_j)$ as the empirical mean and variance of the $\boldsymbol{z}_o$s in this bin.
repeat
    for $i = 1, 2, \cdots, M$ **do**
        Obtain one input observation and its label $(\boldsymbol{x}^i, y^i)$ from training batch.
        Sample $\boldsymbol{z}_o^i \sim q_{\Theta_o}(\boldsymbol{z}_o|\boldsymbol{x}^i)$, $\boldsymbol{z}_p^i \sim q_{\Theta_p}(\boldsymbol{z}_p|\boldsymbol{x}^i)$, $\boldsymbol{z}_c^i \sim q_{\Theta_c}(\boldsymbol{z}_c|\boldsymbol{x}^i)$, $\boldsymbol{z}_s^i \sim q_{\Theta_s}(\boldsymbol{z}_s|\boldsymbol{x}^i)$
        Input $\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s, \boldsymbol{Z}_o = \boldsymbol{z}_p^i, \boldsymbol{z}_c^i, \boldsymbol{z}_s^i, \boldsymbol{z}_o^i$, into the decoder, classifiers and compute $p_\Phi(\boldsymbol{x}^i|\boldsymbol{z}_p^i, \boldsymbol{z}_c^i, \boldsymbol{z}_s^i, \boldsymbol{z}_o^i), p_\Psi(y^i|\boldsymbol{z}_p^i, \boldsymbol{z}_c^i, \boldsymbol{z}_s^i), p_\eta(y^i|\boldsymbol{z}_p^i), p_\delta(z_c^i|y^i, \boldsymbol{z}_s^i)$.
    end for
    Update $\Theta, \Phi, \Psi, \eta, \delta$ by minimizing the training objective in Eq. (5) via gradient descent.
    for $i = 1, 2, \cdots, M$ **do**
        Obtain one input observation $\boldsymbol{x}^i$
        Input $\boldsymbol{X} = \boldsymbol{x}^i$ into the encoder $q_{\Theta_o}(\boldsymbol{Z}_o|\boldsymbol{X})$, obtain $\boldsymbol{z}_o^i = \boldsymbol{\mu}_{\boldsymbol{z}_o}(\boldsymbol{x}^i)$
    end for
    Cluster $\{\boldsymbol{z}_o^i\}_{i=1}^M$ into J bins with Kmeans algorithm. For each bin $U = u_j$, we update the mean and variance for $p(\boldsymbol{Z}_o|U_x = u_j)$ as the empirical mean and variance of the $\boldsymbol{z}_o$s in this bin.
until Converge

Assumptions in the training framework: One of the main challenges in our training framework is the large number of unobserved variables. Out of the 7 variables in the SCM model, only 2 are directly observed. This situation necessitates making numerous assumptions about the distributions associated with $\boldsymbol{Z}_{cmb}, \boldsymbol{Z}_o, U_x$. However, despite these foundational assumptions, our framework consistently showcases its ability to discern and assimilate causal representations, and this is commendable, especially given that it operates within a singular observational training domain and doesn't require any domain-specific knowledge.

Identifiability of $\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s, \boldsymbol{Z}_o$. By adopting a suitable neural network architecture and consistently applying Gaussian assumptions within our learning framework, we can ascertain the identifiability of the determined $\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s, \boldsymbol{Z}_o$ within our model. This is bolstered by sophisticated theoretical findings presented in Kivva et al. [2022]. Comprehensive derivations will be provided in the Appendix C.

4 Experiments

We showcase the proficiency of our suggested approach in out-of-distribution (OOD) prediction tasks. Our approach termed Causal Markov Blanket Representation Learning (CMBRL), is benchmarked against leading domain generalization (DG) methods.

Dataset. We assess our CMBRL approach using the colored-MNIST (CMNIST) dataset [Mao et al., 2022] and PACS [Li et al., 2017] dataset. For the CMNIST dataset, the training domain links digits to pre-determined colors, while in the test domain, digits and colors are independent. We adopt the most challenging setting from [Mao et al., 2022], accentuating the distributional discrepancy and emphasizing spurious color-digit correlations. For PACS, it comprises images from four domains: Photo (P), Art painting (A), Cartoon (C), and Sketch (S). Each domain labels images in 7 categories.

Baselines: For CMNIST, we compare three types of approaches: the correlation-based classifier such as the ERM approach; the causal DG approaches such as IRM[Arjovsky et al., 2019], GenInt[Mao et al., 2021], and CTrans[Mao et al., 2022]; and other non-causal DG methods such as RSC[Huang et al., 2020]. For PACS, We've integrated additional baselines, encompassing causal DG methods like SageNet[Nam et al., 2021] and MatchDG[Mahajan et al., 2021], as well as non-causal DG strategies such as DRO[Sagawa et al., 2019], MLDG[Li et al., 2017], CORAL[Sun and Saenko, 2016], Mixup[Yan et al., 2020], and etc.

Implementation Details. On the CMNIST dataset, we employ a two-layer MLP for all encoders, decoders, and classifiers. For PACS, we use a ResNet-50 pre-trained on ImageNet as the encoder backbone and choose decoders with matching complexity. We set $|U_x| = J = 2$ for CMNIST and $J = 3$ for PACS. Results are averaged over 5 trials.

Table 1: Comparison with SOTA methods on CMNIST

Algorithms	Prediction Acc (%)	
	In-distribution	OOD
ERM	**99.6**	12.3
RSC	96.3	20.5
IRM	98.4	19.9
GenInt	58.5	31.6
CTrans	82.9	51.4
CMBRL	96.3	**61.8**

Table 2: Comparison with SOTA methods on PACS.

Algorithms	PACS				
	A	**C**	**P**	**S**	**Avg**
ERM	84.7±0.4	80.8±0.6	97.2±0.3	79.3±1.0	85.5
GroupDRO	83.5±0.9	79.1±0.6	96.7±0.3	78.3±2.0	84.4
MLDG	85.5±1.4	80.1±1.7	97.4±0.3	76.6±1.1	84.9
CORAL	88.3±0.2	80.0±0.5	97.5±0.3	78.8±1.3	86.2
MMD	86.1±1.4	79.4±0.9	96.6±0.2	76.5±0.5	84.6
RSC	85.4±0.8	79.7±1.8	97.6±0.3	78.2±1.2	85.2
Mixup	86.1±0.5	78.9±0.8	97.6±0.1	75.8±1.8	84.6
DANN	86.4±0.8	77.4±0.8	97.3±0.4	73.5±2.3	83.6
CDANN	84.6±1.8	75.5±0.9	96.8±0.3	73.5±0.6	82.6
MTL	87.5±0.8	77.1±0.7	96.4±0.8	77.3±1.8	84.6
ARM	86.8±0.6	76.8±0.7	97.4±0.3	79.3±1.2	85.1
IRM	84.8±1.3	76.4±1.1	97.2±0.3	79.3±1.0	83.5
SagNet	87.4±1.0	80.7±0.6	97.1±0.1	80.0±0.4	86.3
MatchDG	85.7±1.6	82.5±0.7	**97.9**±0.7	77.3±1.1	85.9
CMBRL	**88.3**±0.3	**84.3**±0.2	96.3±0.0	**81.0**±0.1	**87.5**

Results and Analysis. From Table 1, our CMBRL method significantly surpasses SOTA methods in OOD performance on the CMNIST dataset, while achieving comparable in-distribution accuracy. Our experimental setup on CMNIST treats each color-digit combination as a distinct domain, leading to two primary domains ($|U_x| = 2$), which aligns with our SCM assumptions. Empirically, there's a clear distinction in color information across these domains, as highlighted by the significant disparity between $p(\boldsymbol{z}_o|u_x = 0)$ and $p(\boldsymbol{z}_o|u_x = 1)$. This differentiation aids CMBRL in effectively distinguishing between color ($\boldsymbol{z}_o$) and shape ($\boldsymbol{z}_{cmb}$) features, resulting in enhanced OOD prediction accuracy. According to Table 2, on the real dataset PACS, our method's performance improvement isn't as pronounced as the CMNIST dataset, possibly because of the latent confounder between $\boldsymbol{z}_o$ and y. This situation causes $p(y|\boldsymbol{z}_{cmb})$ to be non-invariant during distributional shifts. We aim to conduct interventional inference in the future to eliminate the influence of this latent confounder on y.

Ablation Study. In Appendix D, we provide ablation studies. These encompass prediction performance based on a subset of CMB representations and an analysis of the hyperparameter J. In conclusion, the CMB features contain the maximum information for prediction. Its classification accuracy notably exceeds that of any subset within the CMB.

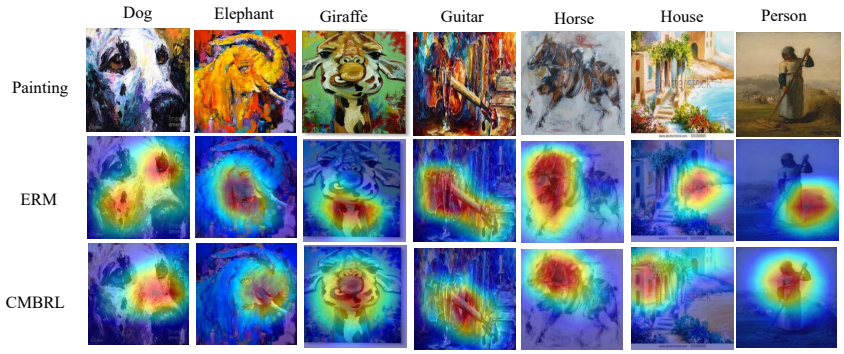

Figure 3: Visualization of CMB representation using GradCAM [Selvaraju et al., 2019] on PACS

5 Conclusion

In this paper, we introduce a novel approach to deep learning by focusing on causal relationships within latent variables, particularly emphasizing the CMB features. These features, by their intrinsic stability and invariance under distribution shifts, hold the promise to revolutionize the way models handle OOD predictions. Our method, rooted in causality, offers a practical and robust solution to enhance model generalization without the need for multi-domain data and domain-specific knowledge. Experimentally, we demonstrate its superiority on the CMNIST dataset and PACS dataset. Further evaluations across diverse scenarios will be our future work.

References

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

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Inf. Proc. Systems*, 2021a.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. Empirical or invariant risk minimization? a sample complexity perspective. In *International Conference on Learning Representations*, 2021b.

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. 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.

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

Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1): 46–100, 2021.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021.

Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, 2022.

Peng Cui, Zheyan Shen, Sheng Li, Liuyi Yao, Yaliang Li, Zhixuan Chu, and Jing Gao. Causal inference meets machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3527–3528, 2020.

Tian Gao, Ziheng Wang, and Qiang Ji. Structured feature selection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4256–4264, 2015.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020.

Dominik Janzing. Causal regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. *arXiv preprint arXiv:2207.01603*, 2022.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020.

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

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18041–18050, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6979–6987, 2017.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2021.

Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531, 2022.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

Jean-Philippe Pellet and André Elisseeff. Using markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9(7), 2008.

J Peters, Peter Buhlmann, and N Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. arxiv. *Methodology*, 2015.

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

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

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL `http://dx.doi.org/10.1007/s11263-019-01228-7`.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *The Journal of Machine Learning Research*, 23(1):10994–11048, 2022.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

Baochen Sun and Kate Saenko. 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*, pages 443–450. Springer, 2016.

Jiachen Sun, Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, Dan Hendrycks, Jihun Hamm, and Z Morley Mao. Certified adversarial defenses meet out-of-distribution corruptions: Benchmarking robustness and simple baselines. *arXiv preprint arXiv:2112.00659*, 2021.

Yuan Tan and Zhifa Liu. Feature selection and prediction with a markov blanket structure learning algorithm. In *BMC bioinformatics*, volume 14, pages 1–3. BioMed Central, 2013.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34:237–250, 2021.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Jiajin Zhang, Hanqing Chao, Xuanang Xu, Chuang Niu, Ge Wang, and Pingkun Yan. Task-oriented low-dose ct image denoising. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 441–450. Springer, 2021.

Jiajin Zhang, Hanqing Chao, Amit Dhurandhar, Pin-Yu Chen, Ali Tajer, Yangyang Xu, and Pingkun Yan. When neural networks fail to generalize? a model sensitivity perspective. In *Proceedings of the 37th AAAI conference on artificial intelligence*, 2023.

A Derivation of Eq. (2)

$$
\begin{aligned}
& p(\boldsymbol{x}, y, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, u_x) \\
=& p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_p)p(\boldsymbol{z}_s)p(y|z_p)p(\boldsymbol{z}_c|y, z_s)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \quad \text{BN chain rule} \\
=& p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_p)p(\boldsymbol{z}_s)p(y|\boldsymbol{z}_p)\frac{p(y|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c|\boldsymbol{z}_s)}{p(y|\boldsymbol{z}_s)}p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o)\,\mathrm{d}\boldsymbol{z} \quad \text{Bayes' theorem} \\
=& p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_p)p(\boldsymbol{z}_s)p(y|\boldsymbol{z}_p)\frac{p(y|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c|\boldsymbol{z}_s)}{p(y)}p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \quad \boldsymbol{Z}_s \perp\!\!\!\perp Y \\
=& p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_s)p(\boldsymbol{z}_p|y)p(y|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c|\boldsymbol{z}_s)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \quad \text{Bayes' theorem: } p(\boldsymbol{z}_p)p(y|\boldsymbol{z}_p) = p(\boldsymbol{z}_p|y)p(y) \\
=& p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_s)p(\boldsymbol{z}_p|y, \boldsymbol{z}_c, \boldsymbol{z}_s)p(y|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c|\boldsymbol{z}_s)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \quad \boldsymbol{Z}_p \perp\!\!\!\perp \boldsymbol{Z}_c, \boldsymbol{Z}_s|Y \\
=& p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_s)p(\boldsymbol{z}_p, y|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c|\boldsymbol{z}_s)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \quad \text{Product rule} \\
=& p(u_x)p(\boldsymbol{z}_o|u_x)p(\boldsymbol{z}_s)p(y|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_p|\boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_c|\boldsymbol{z}_s)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \quad \text{Product rule} \\
=& p(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_o|u_x)p(u_x)p(\boldsymbol{x}|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_o) \quad \text{Product rule}
\end{aligned}
\tag{8}
$$

B Derivation of Eq. (3)

$$
\begin{aligned}
& -\log p(\boldsymbol{x}, y) \\
=& -\log \sum_{u_x} \int_{\boldsymbol{z}=[\boldsymbol{z}_o, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s]} p(\boldsymbol{x}, y, u_x, \boldsymbol{z}_o, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)d\boldsymbol{z} \\
=& -\log \sum_{u_x} \int_{\boldsymbol{z}=[\boldsymbol{z}_o, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s]} p_\Psi(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_o|u_x)p(u_x)p_\Phi(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z} \quad \text{Substitute } Eq.\ (2) \\
=& -\log \int_{\boldsymbol{z}=[\boldsymbol{z}_o, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s]} p_\Psi(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)\Big[\sum_{u_x} p(\boldsymbol{z}_o|u_x)p(u_x)\Big]p_\Phi(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z} \\
=& -\log \int_{\boldsymbol{z}=[\boldsymbol{z}_o, \boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s]} \frac{p_\Psi(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_o)p_\Phi(\boldsymbol{x}|\boldsymbol{z})}{q_\Theta(\boldsymbol{z}|\boldsymbol{x})}q_\Theta(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} \\
=& -\log \mathbb{E}_{q_\Theta(\boldsymbol{z}|\boldsymbol{x})}\Big[\frac{p_\Psi(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)p(\boldsymbol{z}_o)p_\Phi(\boldsymbol{x}|\boldsymbol{z})}{q_\Theta(\boldsymbol{z}|\boldsymbol{x})}\Big] \\
\leq& -\mathbb{E}_{q_\Theta(\boldsymbol{z}|\boldsymbol{x})}\log\Big[p_\Psi(y|\boldsymbol{z}_c, \boldsymbol{z}_p, \boldsymbol{z}_s)p_\Phi(\boldsymbol{x}|\boldsymbol{z})\frac{p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)}{q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_c, \boldsymbol{z}_s, \boldsymbol{z}_p|\boldsymbol{x})}\frac{p(\boldsymbol{z}_o)}{q_{\Theta_o}(\boldsymbol{z}_o|\boldsymbol{x})}\Big] \quad \text{Jensen's inequality} \\
=& -\mathbb{E}_{q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})}[\log p_\Psi(y|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)] - \mathbb{E}_{q_\Theta}[\log p_\Phi(\boldsymbol{x}|\boldsymbol{z})] \\
& -\mathbb{E}_{q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})}[\log \frac{p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)}{q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})}] - \mathbb{E}_{q_{\Theta_o}(\boldsymbol{z}_o|\boldsymbol{x})}[\log \frac{p(\boldsymbol{z}_o)}{q_{\Theta_o}(\boldsymbol{z}_o|\boldsymbol{x})}] \\
=& -\mathbb{E}_{q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})}[\log p_\Psi(y|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)] - \mathbb{E}_{q_\Theta}[\log p_\Phi(\boldsymbol{x}|\boldsymbol{z})] \\
& + KL\Big(q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)\Big) + KL\Big(q_{\Theta_o}(\boldsymbol{z}_o|\boldsymbol{x})||p(\boldsymbol{z}_o)\Big)
\end{aligned}
\tag{9}
$$

We then define our training objective $\mathcal{L}_{nll}$ as:

$$
\begin{aligned}
\mathcal{L}_{nll} \coloneqq& -\mathbb{E}_{q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})}[\log p_\Psi(y|\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)] - \mathbb{E}_{q_\Theta}[\log p_\Phi(\boldsymbol{x}|\boldsymbol{z})] \\
& + KL\Big(q_{\{\Theta_p, \Theta_c, \Theta_s\}}(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s)\Big) + KL\Big(q_{\Theta_o}(\boldsymbol{z}_o|\boldsymbol{x})||p(\boldsymbol{z}_o)\Big)
\end{aligned}
$$

C Identifiability of $\boldsymbol{Z}_p, \boldsymbol{Z}_c, \boldsymbol{Z}_s, \boldsymbol{Z}_o$

Kivva et al. [2022] provides a comprehensive theoretical foundation for the identifiability of the learned variable, $\boldsymbol{Z}$, based on a set of specified assumptions on both the distribution $p(\boldsymbol{z})$ and the mapping function, f, from $\boldsymbol{z}$ to $\boldsymbol{x}$. The key assumptions are:

1. Distribution of $\boldsymbol{z}$: $p(\boldsymbol{z})$ should be a Gaussian mixture model, possibly degenerate, with at least one component ($J \geq 1$).
   In our setup, $p(\boldsymbol{z}_o)$ is designed as a mixture Gaussian distribution, formulated as $p(\boldsymbol{z}_o) = \sum_{j=1}^{J} \lambda_j \mathcal{N}(\mu_j, \Sigma_j)$, with associated probabilities $p(U_x = u_j) = \lambda_j$ and a total of $|U_x| = J$ outcomes. Concurrently, $p(\boldsymbol{z}_{cmb})$ adheres to a standard normal distribution, which also satisfies Assumption 1 .
2. Function f: f should be a piecewise affine function.
   In our implementation, we employ a multilayer perceptron (MLP) with leaky ReLU activations, which naturally aligns with the assumption of a piecewise affine function.
3. Injectivity of f: The function f must be injective.
   According to Corollary H.4 in Kivva et al. [2022], an MLP featuring leaky ReLU activations and an incrementally increasing count of hidden neurons is deemed injective. In our model, we configure the decoder (represented by f) as a leaky ReLU network. In particular, we design the decoder with neural networks with a progressive increasing number of neurons. Hence, our decoder f meets assumptions (2) and (3).

Given the above, we can assertively conclude that the derived $\boldsymbol{Z}$ in our framework is identifiable up to an affine transformation.

D Ablation Study

D.1 Prediction performance of various causal representations

Table 3: Comparison between various causal representations on CMNIST data set.

Algorithms	Prediction Acc (%)	
	In-distribution	OOD
ERM	**99.6**	12.3
Prediction with $\boldsymbol{Z}_p$	91.8	37.0
Prediction with $\boldsymbol{Z}_c$	58.7	43.0
Prediction with $\boldsymbol{Z}_s$	46.9	27.3
Prediction with $\boldsymbol{Z}_p, \boldsymbol{Z}_c$	93.1	48.8
Prediction with $\boldsymbol{Z}_p, \boldsymbol{Z}_s$	92.2	32.5
Prediction with $\boldsymbol{Z}_c, \boldsymbol{Z}_s$	78.3	45.4
CMBRL	96.3	**61.8**

Recent research has extensively debated the choice of causal representations for achieving generalizable prediction. Notably, Lopez-Paz et al. [2017] empirically demonstrated that anti-causal features ($\boldsymbol{z}_c$) of the target yield superior prediction performance under distribution shifts compared to causal features ($\boldsymbol{z}_p$). Many state-of-the-art causal representation learning methods [Mao et al., 2022, Arjovsky et al., 2019], however, predominantly rely on $\boldsymbol{z}_p$ within their assumed SCM.

In light of this, we conduct predictions using $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s$ and compare them to our CMBRL approach, which incorporates all three types of causal representations. Our CMBRL framework involves training classifiers for both $\boldsymbol{z}_{cmb}$ and $\boldsymbol{z}_p$. We retain the trained encoder parameters and focus on training the classifiers $p(y|\boldsymbol{z}_c)$ and $p(y|\boldsymbol{z}_s)$.

As per the empirical results presented in Table 3, all three types of causal representations demonstrate improved OOD prediction performance compared to the correlation-based ERM. However, particularly in the case of our derived $\boldsymbol{z}_c$ and $\boldsymbol{z}_s$, this enhanced OOD performance is accompanied by a significant compromise in in-distribution prediction performance. In contrast, our CMBRL, which harnesses all types of causal representations, achieves superior performance in both in-distribution and OOD scenarios than any subset of CMB representations. This validates our model's assumption that $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s$ are all indispensable and necessary to mitigate the spurious correlations between $\boldsymbol{z}_o$ and the target variable y.

D.2 The number of domains

In this section, we delve into the impact of adjusting the number of domains, denoted as $|U_x|$, on out-of-distribution (OOD) prediction performance. Our approach involves systematically increasing the values of $|U_x|$ from 1 to 5, and the resulting prediction performance is visualized in Figure 4.

The insights gleaned from Figure 4 shed light on our method's performance in different scenarios involving the number of domains, represented as $|U_x|$. Notably, our approach demonstrates its weakest performance when $|U_x|$ is set to 1. In such a configuration, an assumption is made that $\boldsymbol{z}_o$ adheres to a Gaussian distribution, with its prior distribution mirroring that of $\boldsymbol{z}_p, \boldsymbol{z}_c, \boldsymbol{z}_s$. This assumption results in a compromise regarding the asymmetry regularization between the spurious representation $\boldsymbol{z}_o$ and CMB representation $\boldsymbol{z}_{cmb}$ of representations, ultimately leading to suboptimal disentanglement. The rectification of this asymmetry issue becomes apparent when $|U_x|$ is increased to a value greater than or equal to 2. However, the performance outcomes tend to remain relatively consistent for cases where $|U_x|$ exceeds 2. The selection of the optimal $|U_x|$ value depends on the disparities inherent in the true data distributions across each domain. In cases involving observational datasets that lack domain-specific information, opting for $|U_x| = 2$ can still yield a reasonably well-disentangled set of representations.

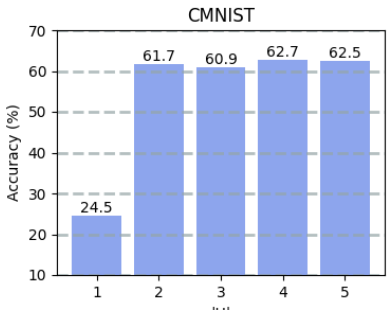

Figure 4: Ablation study on the influence from the number of domains $|U_x|$.