Towards Cross-Modal Causal Structure and Representation Learning

Haiyi Mao HAM112@PITT.EDU
Joint CMU-Pitt PhD Program in Computational Biology, University of Pittsburgh

Hongfu Liu HONGFULIU@BRANDEIS.EDU
Michtom School of Computer Science, Brandeis University

Jason Xiaotian Dou JASON.DOU@PITT.EDU
Department of Electrical and Computer Engineering, University of Pittsburgh

Panayiotis V. Benos PBENOS@UFL.EDU
Department of Epidemiology, University of Florida
Joint CMU-Pitt PhD Program in Computational Biology, University of Pittsburgh

Abstract

Does the SARS-CoV-2 virus cause patients' chest X-Rays ground-glass opacities? Does an IDH-mutation cause differences in patients' MRI images? Conventional causal discovery algorithms, although well developed to uncover the cause-effect relationships on structured data, cannot elucidate causal relations between unstructured images and structured scalar variables due to the complexity of the former. In this paper, we consider causal discovery between images and structured (scalar) variables. Specifically, we derive low dimensional image representations to analyze with structured variables. We propose a two-module amortized variational algorithm named **C**ross-**M**odal Variational **C**ausal representation and structure **L**earning (*CMCL*). *CMCL* jointly learns identifiable representations given a set of independent structured variables and causal relations via formulating latent representations and structured variables into a direct acyclic graph. Moreover, we further enforce counterfactual invariance/variance onto representations. We demonstrate that *CMCL* outperforms other related methods on synthetic datasets and validate causal relations on semi-synthetic datasets by visualization.

Keywords: Causal Structure Learning, Causal Discovery, Multi-Modality, Medical Image Analysis.

1. Introduction

Finding underlying causal relations is a fundamental task in many disciplines, including economics (Wager and Athey, 2018; Athey and Imbens, 2006), biology (Graham et al., 2021), and medicine (Raghu et al., 2018). Causal discovery or causal structure learning aims to recover cause-effect relations by utilizing statistical properties of observational data. Tremendous efforts have been made to develop causal discovery algorithms on observational structured data. These algorithms can be divided into three categories. One prominent category is constraint-based methods which, under appropriate assumptions, recover the underlying causal structure based on conditional independence relationships of the variables (Spirtes and Zhang, 2016; Spirtes et al., 2000). The second category is score-based methods (Chickering and Heckerman, 1997; Chickering, 2002; Huang

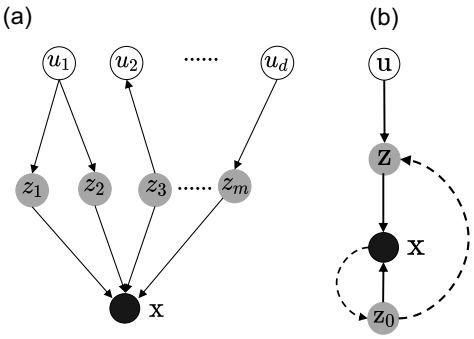

Figure 1: (a) Illustration of data generating processes. Latent variables **z** generate images **x** which are causally related to structured variables **u**. (b) Demonstration of the variational inference variables architecture. The solid lines indicate generative models. The dash lines denote variational approximation. Note here, the causal data generating processes are different from generative models. So $\mathbf{u} \rightarrow \mathbf{z}$ (b) is not equal to all the variables of **u** that causes **z**.

et al., 2018) which optimize a score function that characterizes conditional independence. The last category is based on functional causal models which infer the causal direction by exploiting the independence between the noise and the cause (Zhang and Hyvärinen, 2009; Hoyer et al., 2008; Zhang and Hyvarinen, 2012; Shimizu et al., 2006). With increasing data volumes and modalities (Liang et al., 2020; Boehm and et al., 2021), there is now an urgent need to extend such algorithms to unstructured data so questions like these can be addressed:

Example 1 Does COVID-19 cause symptoms on chest X-Rays?

Example 2 Can we observe any phenotypes on MRI images that are caused by gene IDH mutations (Cluceru and et al., 2021)?

One intuitive solution is to first extract the latent representations[1] from images and then apply the conventional algorithms for causal discovery on these and the structured data (Mao et al., 2022b; Dou et al., 2022a,b,c). However, when the image representation is extracted independently in existing deep learning models, they might not be causally linked to the structured variables; this impairs the downstream causal discovery due to high dimensionality or noisiness. Moreover, the representations, typically learned by neural networks, tend to be highly correlated, which makes it prohibitive for causal discovery, as conditional independence tests on correlated variables are problematic (Spirtes and Zhang, 2016). Although there are recent advances on unsupervised disentangled representation learning with the independence enforced (Higgins et al., 2017; Kim and Mnih, 2018; Mathieu et al., 2016; Chen et al., 2018), they have been proved to be non-identifiable (Locatello et al., 2019) which makes downstream causal discovery unfeasible. In addition, in such approaches, the counterfactual variance/invariance of the representations and their causal relations is not ensured.

Contributions. In this paper, we develop a new algorithm, **C**ross-**M**odal Variational **C**ausal representation and structure **L**earning (*CMCL*), for discovering possible cause-effect relations among structured and unstructured data modalities. Our major

1. In this paper, we use latent representations and latent variables interchangeably.

contributions are: **(1)** We consider a new problem on discovering the causal relations for multi-modal structured variables (continuous, categorical). To the best of our knowledge, this problem has not been addressed yet. **(2)** We propose a model *CMCL* to jointly learn the causal representations and causal structure cross variable modalities by amortized variational inference, which leads to *identifiable*, *interpretable* and *causal* related representations. **(3)** We introduce a counterfactual regularizer based on contrastive learning to enforce the counterfactual invariance/variance. **(4)** We present the theoretical conditions of the model's identifiability and convergence. **(5)** We demonstrate *CMCL* outperforms other combinations of representation learning and causal discovery methods.

2. Problem Definition

The cross-modal data in our research problem consists of unstructured image and structured (continuous, categorical) variables. For l-th data sample, $1 \leq l \leq n$ (n is the number of data samples), $\mathbf{x}^{(l)}$ and $\mathbf{u}^{(l)}$ denote the image (e.g., chest X-Ray) and d-dimensional structured variable vector (e.g., BMI, age, sex, and gene mutant status), respectively. For notational convenience, we use $\mathbf{x}, \mathbf{u}$ to denote the sample without the data point index, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{u} = [u_1, u_2, \dots u_d] \in \mathbb{R}^d$. To better understand the causal relationships between the unstructured $\mathbf{x}$ and structure $\mathbf{u}$, in this section, we elaborate our assumptions and formulate the problem definition.

2.1. Assumptions

To reveal the causal relationships between the unstructured $\mathbf{x}$ and structured $\mathbf{u}$, it is necessary to elaborate the image data generating processes. We hold an assumption on image data generation shown in Fig. 1(a), that the image object $\mathbf{x} \in \mathcal{X}$ is generated by a set of latent variables $\mathbf{z} = [z_1, z_2, \dots, z_m] \in \mathbb{R}^m$. Furthermore, $\mathbf{g} \in \mathcal{G}$ is the image generation function $\mathbb{R}^m \to \mathcal{X}$ such that $\mathbf{g}(\mathbf{z}) + \mathbf{e} = \mathbf{x}$, where $\mathbf{e}$ is the noise term with a probability density of $p_{\mathbf{e}}(\mathbf{x} - \mathbf{g}(\mathbf{z}))$. Note, we use latent representations and latent variables interchangeably. Formally, our first image data generation is as follows:

Assumption 1:
(Image Generating Processes) We assume that the images $\mathbf{x}$ are generated by a set of latent variables $\mathbf{z}$ via an injective generating function $\mathbf{g}$ such that

$$\mathbf{g}(\mathbf{z}) + \mathbf{e} = \mathbf{x}. \tag{1}$$

$\{\mathbf{x} \in \mathcal{X} | \mathcal{C}_{\mathbf{e}}(\mathbf{x}) = 0)\}$ has measure zero, where $\mathcal{C}_{\mathbf{e}}$ is the characteristic function of $p_{\mathbf{e}}$.

Although this assumption on image data generating processes is similar to Wang and Jordan (2021) , in our framework, we further assume that there exists causal relationships between the latent variables $\mathbf{z}$ and $\mathbf{u}$. In addition, we distinguish the image generation relations from the causal relations. To be specific, the causal assumptions are as follows:

Assumption 2:
(Causal Assumptions) Given a DAG as a tuple $(V, E, \mathbf{f})$, where $V = \mathbf{z} \cup \mathbf{u}$; E is the set of edges in the causal Directed Acyclic Graph (DAG); $\mathbf{f} = [f_1, f_2, \dots f_d]$ is a set of non-linear causal functions corresponding to structured variables $\mathbf{u} = [u_1, u_2, \dots, u_d]$, then we have

1. Causal: $u_i \to z_j \Rightarrow f_j(u_i) + \epsilon_j = z_j$;
2. Anti-causal: $u_i \leftarrow z_j \Rightarrow f_i(z_j) + \epsilon_i = u_i$;
3. For all $f_i \in \mathbf{f}$ are non-linear and three-times differentiable;
4. There is no confounder between $\mathbf{z}$ and $\mathbf{u}$;

5. The structured variables u_i are mutually independent,

where $\rightarrow$ means a causal direction, $\Rightarrow$ indicates the corresponding causal functions.

We use "causal" and "anti-causal" to denote two opposite causal directions between $\mathbf{u}$ and $\mathbf{z}$ as Veitch et al. (2021). To better explain Assumption 2, here are some examples of the causal/anti-causal relations: (1) COVID-19 causes the ground-glass opacity of the chest X-Ray, (2) The latent variables encoding the lung air density cause the lung volume ratio to be less than a given threshold (-856 HU) of the CT scan (Lynch and Al-Qaisi, 2013). The fact that causal functions are non-linear additive models, guarantees the identifiability of the causal relations (Hoyer et al., 2008; Zhang and Hyvarinen, 2012). Moreover, for simplicity, we assume the causally related representation to u_i is a 1-dimensional scalar. However, it can be easily extended to higher dimensions. It's worth noting that *CMCL* is different from previous work (Yang et al., 2021) in problem settings and assumptions. CausalVAE is to discover the causal relations among latent variables; *CMCL* aims to identify relations between latent and external(scalar) variables. CausalVAE also implicitly assumes the latent variables have the same causal structures as scalar(external) variables, but *CMCL* does not. Additionally, we make the assumption (5) of mutual independence due to the instability of variational inference (Aneja et al., 2021). Via these causal assumptions, the latent representations $\mathbf{z}$ entail conditional independence given $\mathbf{u}$. Thus, by utilizing this property from iVAE (Khemakhem et al., 2020), we can learn identifiable latent representations. Formally, we have the following conditional independent prior assumption,

Assumption 3:
(Conditional priors) The conditional prior $p(\mathbf{z}|\mathbf{u})$ follows exponential distribution as below:

$$p_{\boldsymbol{T},\boldsymbol{\eta}}(\mathbf{z}|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\mathbf{u})} \exp[\sum_{j=1}^{k} \boldsymbol{T}_{i,j}(z_i)\boldsymbol{\eta}_{i,j}(\mathbf{u})], \tag{2}$$

where z_i is the coordinate index in latent variables vector $\mathbf{z}$, Q_i is the base measure, $Z_i(\mathbf{u})$ is the normalizing constant, $\boldsymbol{T}_i = (\boldsymbol{T}_{i,1}, \ldots, \boldsymbol{T}_{i,k})$ is the sufficient statistics and $\boldsymbol{\eta}_i(\mathbf{u}) = (\boldsymbol{\eta}_{i,1}(\mathbf{u}), \ldots \boldsymbol{\eta}_{i,k}(\mathbf{u}))$ is the corresponding parameters depending on $\mathbf{u}$. This conditional prior is similar to iVAE (Khemakhem et al., 2020), which shows that with the auxiliary variables $\mathbf{u}$ and corresponding exponential conditional distribution, the disentangled representations $\mathbf{z}$ are identifiable. In this paper, we show under the conditional prior assumption, the identifiability of $\mathbf{z}$ and $\mathbf{g}$ can still be ensured in our relatively complicated scenario. We give the formal definition of identifiability in Appendix C.

2.2. Problem Statement

Since we aim to discover the causal structures between latent $\mathbf{z}$ and $\mathbf{u}$ where $\mathbf{z}$ are subject to data generating processes $\mathbf{g}(\mathbf{z}) + \mathbf{e} = \mathbf{x}$, we can formulate this problem as follows:

$$\begin{aligned} &\arg\min_{\mathbf{f}} \sum_{i=1}^{d+m} \ell(v_i - f_i(Pa(v_i))), \\ &\text{subject to } \mathbf{g}(\mathbf{z}) + \mathbf{e} = \mathbf{x}, v_i \in \{\mathbf{u} \cup \mathbf{z}\}, f_i \in \mathbf{f}, \end{aligned} \tag{3}$$

where ℓ is the loss function.

Different from the conventional causal relation discovery methods (Spirtes and Zhang, 2016; Chickering and Heckerman, 1997; Spirtes and Zhang, 2016), there are three main challenges to the above problem. (1) How to learn a generation function $\mathbf{g}$ and corresponding representations $\mathbf{z}$ which are identifiable and amenable to causal discovery. Previous works usually adopt a conditional VAE to infer $\mathbf{g}$ and $\mathbf{z}$ by an approximated posterior $p(\mathbf{z}|\mathbf{u}, \mathbf{x})$. Unfortunately,

these methods flatten $\mathbf{x}$ with $\mathbf{u}$ at the cost of losing spatial information of $\mathbf{x}$, which is crucial for large medical images. (2) How can we learn a causal DAG consisting of representations of unstructured data and structured variables? Traditional causal DAG learning methods struggle to directly apply to image features due to their prohibitively high-dimensions. (Pearl, 2009; Spirtes and Zhang, 2016). (3) There are some recent works on counterfactual variance (Wang and Jordan, 2021; Mitrovic et al., 2020), but none of them are under a causal discovery framework. Hence, the third challenge is how to ensure that the counterfactual variance/invariance in the representations learning processes is suitable for causal discovery tasks. To address these challenges, we propose a model with 3 components. **First** is the two-module representation learning component, in the module one we learn low dimensional embeddings of images; in module 2 the causal representation $\mathbf{z}$ is learned based on latent embeddings and $\mathbf{u}$. **Second** is the causal DAG component, we use the continuous constrained structure learning strategy to recover the causal DAG jointly with $\mathbf{z}$. In the **third** component, we introduce a contrastive counterfactual regularizer to enhance the counterfactual variance/invariance. The main notations are in Table 3.

3. Model

Our goal is to estimate causal representation $\mathbf{z}$, the associated generation function $\mathbf{g}$, and the corresponding causal DAG. To address the challenges mentioned in the previous section, we introduce our Cross-Modal Variational Causal representations and structures Learning (*CMCL*) framework, constituted by a two-module representation learning component, a causal DAG learning component and a counterfactual representations contrastive regularizer. In the two-module representation learning component, the first module aims to learn lower dimensional embeddings to tackle the heterogeneous modality of variables, while the second module is to learn the relevant causal representations. Then the causal DAG learning component seeks to recover the underlying DAG consisting of $\mathbf{u}$ and $\mathbf{z}$. Furthermore, the counterfactual contrastive regularizer strives to enforce the counterfactual variance/invariance to further improve the learning processes. Moreover, as a benefit of *CMCL*, we present the visualization of intervention over the learned causal relations. We will discuss each component respectively in this section. The architecture of model is in Appendix D.2.

3.1. Two-Module Representation Learning

It is intuitive to employ the conditional variational auto-encoder (Sohn et al., 2015) by flattening $\mathbf{x}$ to concatenate with $\mathbf{u}$ to estimate posterior $p(\mathbf{z}|\mathbf{x},\mathbf{u})$. However, unlike previous works (Khemakhem et al., 2020; Yang et al., 2021), considering the heterogeneous structures of multi-modal data, we propose a two-module representation learning model based on VAEs (Kingma et al., 2014; Hassanpour and Greiner, 2021). Specifically, the first module learns low dimension embeddings $\mathbf{z_0}$ which preserve essential information of $\mathbf{x}$. The second module is to estimate causal representation $\mathbf{z}$ by amortized variational inference given $\mathbf{z_0}$ and $\mathbf{u}$. Moreover, $\mathbf{z}$ are causally related to $\mathbf{u}$ such that constitute a causal DAG. The generative model structure shows in Fig. 1(b). Therefore, we have the following conditional generative model given a sample and its independent copy $\mathbf{x}, \tilde{\mathbf{x}}$, we have following

$$\begin{aligned} p_\theta(\mathbf{x},\tilde{\mathbf{x}},\mathbf{z},\mathbf{z_0}|\mathbf{u}) &= p_\psi(\tilde{\mathbf{x}},\mathbf{z_0})p_\xi(\mathbf{x},\mathbf{z}|\mathbf{u}) \\ &= \underbrace{p_\psi(\mathbf{z_0})p_\psi(\tilde{\mathbf{x}}|\mathbf{z_0})}_{\text{latent embeddings}}\ \underbrace{p_\xi(\mathbf{x}|\mathbf{z},\mathbf{u})p_\xi(\mathbf{z}|\mathbf{u})}_{\text{causal representations}}, \end{aligned} \tag{4}$$

where $\mathbf{z}$ and $\mathbf{u}$ are the nodes of a DAG. For simplicity, we denote $\mathbf{x}$ and $\tilde{\mathbf{x}}$ both as $\mathbf{x}$.

According to generative model Fig. 1(b), we have the first equation in Eq. (4), where $p_\psi(\mathbf{x}, \mathbf{z_0}) = p_\psi(\mathbf{z_0})p_\psi(\mathbf{x}|\mathbf{z_0})$ posits the latent embedding module. In addition, $p_\xi(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_\xi(\mathbf{x}|\mathbf{z}, \mathbf{u})p_\xi(\mathbf{z}|\mathbf{u})$ serves the causal representations learning module. In this paper, we use $q_\phi(\mathbf{z}|\mathbf{z_0}, \mathbf{u})$ to approximate the $p_\xi(\mathbf{z}|\mathbf{u}, \mathbf{x})$ instead of $p(\mathbf{z}|\mathbf{x}, \mathbf{u})$ by flattening $\mathbf{x}$ and concatenating with $\mathbf{u}$ to estimate $\mathbf{z}$ directly, as in Khemakhem et al. (2020).

Latent Embedding Module. This module is to learn low dimension embeddings $\mathbf{z_0}$ that preserve sufficient information of $\mathbf{x}$. The problem can be characterized as a generative model $p_\psi(\mathbf{x}, \mathbf{z_0}) = p_\psi(\mathbf{x}|\mathbf{z_0})p_\psi(\mathbf{z_0})$. We treat it as a traditional vanilla VAE, implemented with a convolutional neural network, $\log p_\psi(\mathbf{x})$, as follows:

$$\begin{aligned}\log p_\psi(\mathbf{x}) &= \log \int p_\psi(\mathbf{x}|\mathbf{z_0})p_\psi(\mathbf{z_0})d\mathbf{z_0} \\ &\geqslant \mathbb{E}_{q_\phi(\mathbf{z_0}|\mathbf{x})}[\log p_\psi(\mathbf{x}|\mathbf{z_0}) \\ &- D_{KL}(p_\psi(\mathbf{z_0})||q_\phi(\mathbf{z_0}|\mathbf{x}))],\end{aligned} \tag{5}$$

where the above inequation presents the evidence lower bound (ELBO) (Kingma and Welling, 2014). Thus, the objective of latent embedding module is,

$$\begin{aligned}\mathcal{L}_1 =& \mathbb{E}_{q_\phi(\mathbf{z_0}|\mathbf{x})}[\log p_\psi(\mathbf{x}|\mathbf{z_0}) \\ &- D_{KL}(p_\psi(\mathbf{z_0})||q_\phi(\mathbf{z_0}|\mathbf{x}))].\end{aligned} \tag{6}$$

Causal Representation Learning Module. This module seeks to learn the causal representation $\mathbf{z}$ and generating function $\mathbf{g}$. As shown in Eq. (4), we characterize this problem into a conditional generative model $p_\xi(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_\xi(\mathbf{x}|\mathbf{z}, \mathbf{u})p_\xi(\mathbf{z}|\mathbf{u})$, where $\xi = \{\mathbf{g}, \boldsymbol{T}, \boldsymbol{\eta}\}$ is the domain of parameters describing in Eq. (1) & (2). Following is the $\log p_\xi(\mathbf{x}|\mathbf{u})$ with $\mathbf{z}$ and parameters ϕ,

$$\begin{aligned}\log p_\xi(\mathbf{x}|\mathbf{u}) &= \log \int p_\xi(\mathbf{x}|\mathbf{z}, \mathbf{u})p_\xi(\mathbf{z}|\mathbf{u})d\mathbf{z} \\ &\overset{(1)}{=} \log \int p_\xi(\mathbf{x}|\mathbf{z})p_\xi(\mathbf{z}|\mathbf{u})d\mathbf{z} \\ &\geqslant \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{z_0},\mathbf{u})}[\log(\frac{p_\xi(\mathbf{x}|\mathbf{z})p_\xi(\mathbf{z}|\mathbf{u})}{q_\phi(\mathbf{z}|\mathbf{z_0}, \mathbf{u})})],\end{aligned} \tag{7}$$

where the inequation is the ELBO of the log-likelihood (Khemakhem et al., 2020; Kingma and Welling, 2014). Eq. (7) is a variant of conditional VAE, where the difference is that instead of employing $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$ (Khemakhem et al., 2020; Yang et al., 2021), we amortize the estimation by posterior $q_\phi(\mathbf{z}|\mathbf{z_0}, \mathbf{u})$ to address the heterogeneous structures of multi-modal variables. The second equation (1) holds because $\mathbf{x} \perp\!\!\!\perp \mathbf{u}|\mathbf{z} \Rightarrow p_\xi(\mathbf{x}|\mathbf{z}, \mathbf{u}) = p_\xi(\mathbf{x}|\mathbf{z})$ (see Figure 1(b)). Moreover, given the mutual independence $u_i \perp\!\!\!\perp u_j \in \mathbf{u}$, we can factorize the priors into multiple "sub-priors" to facilitate the disentanglement of representations. We further show the derivation in Appendix C.1.

3.2. Causal DAG Learning

With learned underlying causal representations $\mathbf{z}$, we aim to recover the causal DAG. We follow the previous works (Zheng et al., 2018, 2020) to cast our learning problems into the following format,

$$\min_{W \in \mathbb{R}^{d\times d}} \frac{1}{2n}||V - VW||_F^2 + \lambda||W||_1, \text{ s.t. } h(W) = 0, \tag{8}$$

where $V \in \mathbb{R}^d$ is the node vectors, in our case, nodes $[\mathbf{u}, \mathbf{z}] = V$. W is a weighted adjacent matrix, $h(W) = tr(e^{W\circ W}) - d$, which characterizes the "DAGness". Furthermore the corresponding $\mathbf{f}$, where $v_i = f_i(Pa(v_i)) + \epsilon_i$ can be modeled as a Multi-Layer Perceptrons (MLP), as well as the loss functions can be extended to different link functions such as logistic regression (Zheng et al., 2020). Importantly, we are able to extract the weighted adjacent matrix W from the first layer of MLP denoted as $A^{(1)}$. Specifically, consider an MLP with h hidden layers, a single activation σ: $\mathbb{R} \to \mathbb{R}$, and a input vector $V \in \mathbb{R}^d$ given by

$$\text{MLP}(V; A^{(1)} \dots A^{(h)}) = \sigma(A^{(h)}\sigma(\dots(A^{(1)}(V))), \tag{9}$$

where if the k-th column of $A^{(1)}$ are all zeros, then the output of $\text{MLP}(V; A^{(1)}, \dots A^{(h)})$ will be independent of V i-th coordinate

v_i (Zheng et al., 2020). Thus we have $W_{kj} = 0$ if the k-th column of $A_j^{(1)}$ are all zeros, where $A_j^{(1)}$ denotes the j-th MLP. The DAG learning objective can be formalized as follows:

$$\mathcal{L}_{DAG} = \frac{1}{n}\sum_{i=1}^{d+m} \ell(v_i, \text{MLP}(V)) + \alpha_1 ||A_i^{(1)}||_1 + \alpha_2 h(W), \tag{10}$$

where $v_i \in \{\mathbf{u} \cup \mathbf{z}\}$. Thus, the causal DAG learning problem is transformed into a continuous optimization problem with a DAG-ness constraint. This enables us to learn the latent representation $\mathbf{z}$ and causal DAG jointly.

3.3. Counterfactual Contrastive Regularizer

This constraint leverages a set of counterfactual variance and invariance in the data generation process to improve the learning. By assumption, we have k-th sample pair $(\mathbf{x}^{(k)}, \mathbf{u}^{(k)})$, and corresponding latent representations $\mathbf{z}^{(k)}$. With the causal DAG in Eq. (10), for instance, (1)$u_i \to z_j$, (2)$z_m \to u_n$, then following equations hold,

$$\begin{aligned} &\mathbf{x}^{(k)}(u_i = u_i^{(k)}) \neq \mathbf{x}^{(k)}(u_i \neq u_i^{(k)}), \\ &\mathbf{x}^{(k)}(u_n = u_n^{(k)}) = \mathbf{x}^{(k)}(u_n \neq u_n^{(k)}). \end{aligned} \tag{11}$$

It mainly describes the essential property of two causal structures: (**1**) for *causal* direction, $\mathbf{x}$ could change if u_i changes, (**2**) for *anti-causal* direction, $\mathbf{x}$ should stay the same if u_n changes. Furthermore, we can use the causal representation learning network to enforce above equality and inequality algorithmically, so that Eq. (11) can be formalized into the following objective:

$$\begin{aligned} &\arg\max_p (\\ &\sum_{i\in\Omega^+} p(\tilde{\mathbf{x}}^{(k)}(u_i \neq u_i^{(k)}) \neq \mathbf{x}^k | \mathbf{x} = \mathbf{x}^{(k)} u_i = u_i^{(k)}) - \\ &\sum_{n\in\Omega^-} p(\tilde{\mathbf{x}}^{(k)}(u_n \neq u_n^{(k)}) \neq \mathbf{x}^k | \mathbf{x} = \mathbf{x}^{(k)} u_n = u_n^{(k)})), \end{aligned} \tag{12}$$

where $\tilde{\mathbf{x}}$ is the empirical estimation of image $\mathbf{x}$; $\Omega^+ = \{u_i | u_i \in Pa(z_j), z_j \in \mathbf{z}, u_i \in \mathbf{u}\}$, $\Omega^- = \{u_n | u_n \in Ch(z_m), z_m \in \mathbf{z}, u_n \in \mathbf{u}\}$ $Ch(\cdot)$ is the children set of a variable, $Pa(\cdot)$ is the parents set of a variable in DAG. Instead of discriminating the counterfactual images with causal and anti-causal directions (Yue et al., 2021), we want to distinguish counterfactual images' representations. To begin with, we introduce the definition of the counterfactual representations,

Definition 1 *We define* $\tilde{\mathbf{z}}^{(k)}(u_i \neq u_i^{(k)})$ *as the empirical estimation of counterfactual representation of* $\tilde{\mathbf{x}}^{(k)}(u_i \neq u_i^{(k)})$, *where* $\tilde{\mathbf{z}}^{(k)}(u_i \neq u_i^{(k)}) = \tilde{\mathbf{g}}^{-1}(\tilde{\mathbf{x}}^{(k)}(u_i \neq u_i^{(k)}))$.

With this definition of counterfactual representations, the discrimination problem can be transformed into a contrastive learning task (Yue et al., 2021; Aneja et al., 2021; Bardes et al., 2021) problem. One benefit is that we can utilize the learnt encoding system q_ϕ to save parameter space. In summary, given counterfactual representations $\tilde{\mathbf{z}}^{(k)}(u_i \neq u_i^{(k)})$, we want to train a classifier $\mathbf{D}$ to distinguish the causal counterfactuals $\tilde{\mathbf{z}}^{(k)}(u_i \neq u_i^{(k)}), u_i \in \Omega^+$ and anti-causal counterfactuals $\tilde{\mathbf{z}}^{(k)}(u_n \neq u_n^{(k)}), u_n \in \Omega^-$. Formally, we propose the ***C**ounterfactual **R**epresentation **C**ontrastive **R**egularizer* (*CR*),

$$\begin{aligned} \mathcal{L}_{CR} = \mathbb{E}_{q_\phi}[&\mathbb{E}_{\Omega^+}(\mathbf{D}(\tilde{\mathbf{z}}^{(k)}(u_i \neq u_i^{(k)}))) \\ &+ \mathbb{E}_{\Omega^-}(1 - \mathbf{D}(\tilde{\mathbf{z}}^{(k)}(u_n \neq u_n^{(k)})))]. \end{aligned} \tag{13}$$

CR distinguishes the causal and anti-causal representations to improve the counterfactual variance/invariance of the representation and causal DAG learning with a classifier $\mathbf{D}$.

Overall Objectives. Given the previous ELBOs of two modules, objectives of the causal DAG, and *CR*, we can finalize our objective function as follows:

$$\mathcal{L} = -\mathcal{L}_1 - \mathcal{L}_2 + \mathcal{L}_{DAG} - \mathcal{L}_{CR}. \tag{14}$$

The procedure of *CMCL* is in Algorithm 1.

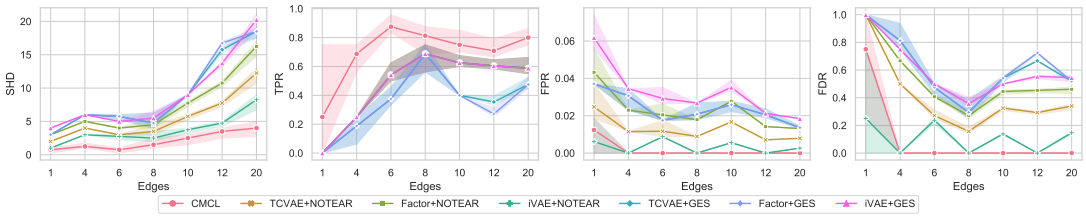

Figure 2: The quantitative evaluation of causal DAG learning. This figure shows the results of the evaluations with structural hamming distance (SHD), true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR). The X-axis is the edge numbers in corresponding simulated DAGs. For notation convenience, we use GES for GS-GES (Huang et al., 2018), Factor for FactorVAE (Kim and Mnih, 2018).

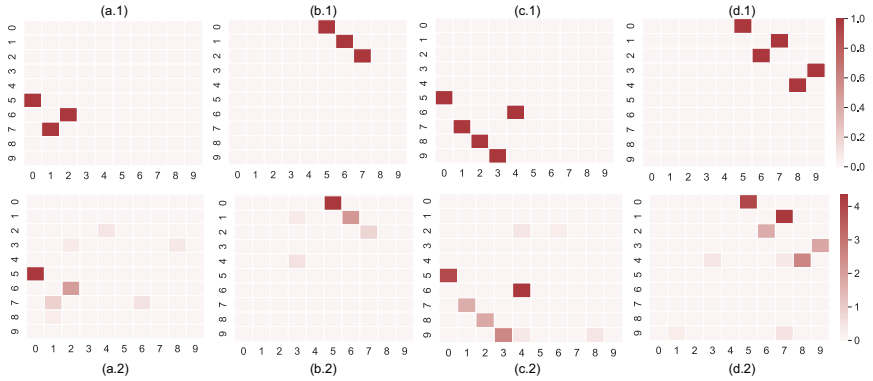

Figure 3: The recovered weighted adjacent matrix compared to the ground truth. This figure shows the ground truth adjacent matrix of W (first row)and the estimated adjacent matrix $\hat{W}$ (second row). There are 5 latent variables (first 5 variables) and 5 structured variables (last 5 variables). (a.1) is the ground truth weighted adjacent matrix $z_1 \to u_1, z_2 \to u_3, z_3 \to u_2$, (a.2) is the estimated weight adjacent matrix. 2.(b.1) is the ground truth weight matrix $u_1 \to z_1, u_2 \to z_2, u_3 \to z_3$, (b.2) is the estimated weight matrix. 3.(c.1) is the ground truth weight matrix $z_1 \to u_1, z_3 \to u_2, z_4 \to u_3, z_5 \to u_4, z_2 \to u_5$, (c.2) is the estimated weight matrix. 3.(d.1) is the ground truth weight matrix $u_1 \to z_1, u_2 \to z_3, u_3 \to z_2, u_4 \to z_5, u_5 \to z_4$, (d.2) is the estimated weight matrix.

4. Experiments

We conduct comprehensive and extensive experiments on synthetic data, semi-synthetic data, real datasets, and corresponding ablation study. Given that underlying ground truth of causal models in real data is usually not known, we generated a series of synthetic datasets to quantitatively evaluate our method. Under different settings, we compare our method with related works in terms of DAG recovery(Figure 2, Figure 3)and representation learning(Table 1). Then we utilize semi-synthesized data to demonstrate that our model can correctly identify the causal relations and further visualize the interventions(Figure 4). Last we use an exist-

ing X-Ray dataset from COVID-19 patients to illustrate our model is able to discover clinical causal relations among cross-modal variables.

4.1. Synthetic Data and Quantitative Results

We use synthetic data to evaluate both learned representation and causal DAGs quantitatively. To our knowledge, there is no method to operate on the same problem setting as *CMCL*, thus, we conduct comparisons with disentangled representation learning methods combined with DAG structure learning methods. Specifically, for representation learning methods we use FactorVAE (Kim and Mnih, 2018), iVAE (Khemakhem et al., 2020), TCVAE (Chen et al., 2018). Regarding causal structural learning algorithms, we choose NOTEAR (Zheng et al., 2020) and GS-GES (Huang et al., 2018). Note that all the compared representation learning methods do not return the ordered representations corresponding to the ground truth $\mathbf{z}$. So we utilize Maximum Information Correlation (MIC) (Kinney and Atwal, 2014) to identify the most correlated coordinate of estimated $\tilde{\mathbf{z}}$ for each ground truth $z_i \in \mathbf{z}$. Then we use the matched $\tilde{\mathbf{z}}$ to evaluate the DAG results. In addition, note GS-GES returns a CPDAG that may contain undirected edges (Huang et al., 2018). In this case, we evaluate them favorably by assuming correct orientation for undirected edges whenever possible as in (Zheng et al., 2018, 2020). We first synthesize $[10, 10, 12, 16, 20, 24, 40]$ variables and $[1, 4, 6, 8, 10, 12, 20]$ edges. The details of synthesizing can be found in Appendix E. We repeat the simulation and experiments 5 times. We adopt commonly used metrics: Structure Hamming Distance (SHD), False Discovery rate (FDR), True Positive Rate (TPR), and False Positive Rate (FPR) (Zheng et al., 2020). The quantitative evaluations of recovered causal DAG are shown in Figure 2. As the figure shows, our method outperforms the compared methods across all measurements. The main reason is that we learn the latent variables and DAGs jointly. For the first dataset, the recovered weighted adjacent matrices are reported in Figure 3. Specifically, the first row shows the corresponding simulated adjacent matrix. The second row illustrates the performance of our model in learning the major causal relations with minor noises. It demonstrates that the conditional prior and jointly learning mechanisms can recover underlying causal DAG successfully

For the perspective of the representation, we use the MIC (Kinney and Atwal, 2014) to evaluate the correlations between learned representations and ground truths. We synthesize $[10, 10, 16, 20]$ variables(nodes) and $[3, 5, 8, 10]$ edges, the mean MIC with standard deviations is reported with simulated DAGs in Table 1. According to the table, our method performs better than all other compared methods in terms of MIC. In summary, the overall excellent performance of *CMCL* is attributed to the joint learning mechanism and CR. Note that iVAE also performs better than other methods, because *CMCL* and iVAE are both weakly-supervised algorithms.

4.2. Semi-Synthetic Data and Intervention Visualization

We demonstrate both causal relations discovery and interventions visualization with MNIST dataset (LeCun and Cortes, 2010). For simplicity, we retrieve 10 different digits images from the dataset. We simulate a structured variable *angle*, then we rotate digits images with $f(angle)$ degrees where $f(\cdot)$ is a non-linear function according to Assumption 2. Moreover, the intervened im-

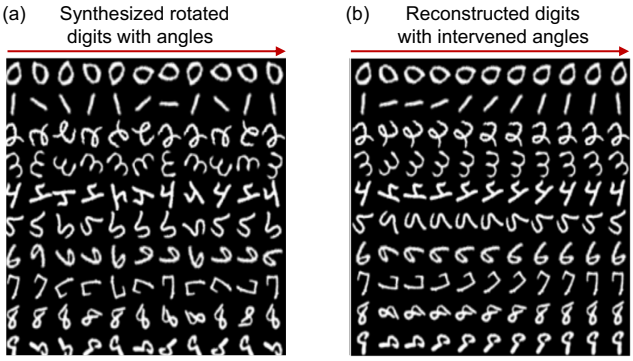

Figure 4: Intervention visualization of rotation. (a) Simulated rotated images by a function of given degree variable $angle \sim Uniform(0, 10)$; (b) We first apply *CMCL* to discover the causal relation *angle* and representation z_1 which encoding the ration feature $angle \to z_1$. Then we intervene *angle* from from 0 to 10 and re-generate images by $p(\mathbf{x}|do(angle))$. The reconstructed images are shown in (b). As expected, the re-generated images rotate correspondingly with intervened *angle* values. Rows are different digits, columns are images intervened with different *angle* values.

Table 1: The MIC comparison among different methods.

Nodes	*CMCL*	βVAE	FactorVAE	iVAE	TCVAE
3/10	$\mathbf{93.5}_{\pm 1.4}$	$33.3_{\pm 5.6}$	$37.9_{\pm 5.8}$	$47.2_{\pm 3.1}$	$40.6_{\pm 5.3}$
5/10	$\mathbf{95.8}_{\pm 1.7}$	$21.7_{\pm 6.8}$	$20.2_{\pm 6.3}$	$48.9_{\pm 4.9}$	$23.4_{\pm 6.5}$
8/16	$\mathbf{88.4}_{\pm 2.5}$	$12.2_{\pm 4.3}$	$13.7_{\pm 4.6}$	$43.6_{\pm 4.2}$	$14.2_{\pm 5.2}$
10/20	$\mathbf{79.3}_{\pm 2.8}$	$12.5_{\pm 4.8}$	$12.6_{\pm 5.2}$	$45.7_{\pm 3.8}$	$13.8_{\pm 5.1}$

ages can be computed as $p(\mathbf{x}|do(u_i = u_i')) \propto p_\theta(\mathbf{x}|\mathbf{z})f_j(u_i = u_i')$ for the causal case 1. The derivation of interventions can be found in Appendix B. Totally, we simulate 1000 samples of images with corresponding *angle*. After training the model, we find out latent variable z_1 is causally related to *angle* as $angle \to z_1$ where z_1 encodes the rotation operation. As expected, we successfully identify the causal relations between *angle* and the rotation operation. In addition, we intervene *angle* by sampling 10 different values from $angle \in [0, 5]$. Next, we use learned relations between *angle* and z_1 to infer the new $\mathbf{z}$ and generate new images. The generated images are shown in Figure 4. From the results, we can see our algorithm correctly recover the causal direction between $angle \to z_1$ where z_1 encodes the rotation operation. And newly generated images are rotated in a corresponding intervened *angle*.

4.3. Experiments on Real Dataset

In this subsection, we evaluate our method on real COVID-19 patients' chest X-ray data (Maguolo and Nanni, 2021).[2] There are 460 COVID-19 patients and 300 non-COVID patients chest X-ray images. We set $\mathbf{u}$ to encode the patient infected with COVID-19 or not. $\mathbf{x}$ are the corresponding chest X-ray images. For the pre-processing, we resize the image to 128×128, and set $\{1, 0\}$ as {COVID, NO-COVID}. We use logistic

2. https://github.com/ieee8023/covid-chestxray-dataset

regression as the link function in the DAG learning component for the $\mathbf{u}$. Then our method returns that $\mathbf{u} \rightarrow z_1$ where z_1 encode the feature of peripheral ground-glass opacity. This is consistent with previous clinical findings (Rousan et al., 2020). Figure 5 shows three representative examples where the heatmap of z_1 represents the peripheral ground-glass opacity. We use GRAD-CAM++ (Chattopadhay et al., 2018) to generate the heatmap. We demonstrate *CMCL* can correctly identify the symptoms.

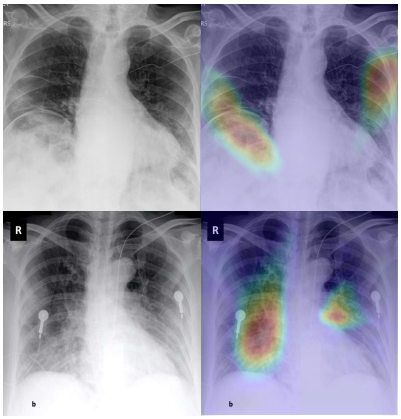

Figure 5: The visualization of the latent representation of the peripheral ground-glass caused by COVID-19.

5. Discussion

In this paper, we address an important causal discovery problem of crossing modalities, from images (unstructured) to scalar (structured) variables. We propose an algorithm, *CMCL*, that is based on a data generating process with structured variables. This algorithm allows the causal relations to be represented by a DAG. The latent causal representations and the DAG are jointly learned in an amortized variational inference manner. Moreover, with our designed regularizer *CR*, the causal relations are further enhanced via the counterfactual variance/invariance. We then discuss the identifiability of latent representations. Under various experiments settings, we show that our method outperforms existing methods, which is attributed to the jointly learning of latent variables and causal relations. The main limitation of this paper is the assumption of bipartite causal graphs in Assumption 2. However, this assumption allows the latent representations to be identifiable, without needing additional assumptions and information. In future work, we would like to adopt to more complicated data generating processes with considering confounders.

Acknowledgments

This work is supported by NIH R01 HL1 59805, R01 HL157879, R01 AA028436.

References

Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. In *NeurIPS*, volume 34, 2021.

Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Kevin M Boehm and et al. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, pages 1–13, 2021.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational

autoencoder: Learning disentangled representations from grouped observations. In *AAAI*, 2018.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv:1802.04942*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016.

David Maxwell Chickering. Optimal structure identification with greedy search. *JMLR*, 3(Nov):507–554, 2002.

David Maxwell Chickering and David Heckerman. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine learning*, 29(2):181–212, 1997.

Julia Cluceru and et al. Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. *Neuro-Oncology*, 10 2021. ISSN 1522-8517. noab238.

Jason Xiaotian Dou, Minxue Jia, Nika Zaslavsky, Runxue Bao, Shiyi Zhang, Ke Ni, Paul Pu Liang, Haiyi Mao, and Zhihong Mao. Learning more effective cell representations efficiently. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022a.

Jason Xiaotian Dou, Lei Luo, and Raymond Mingrui Yang. An optimal transport approach to deep metric learning (student abstract). In *AAAI*, 2022b.

Jason Xiaotian Dou, Alvin Qingkai Pan, Runxue Bao, Haiyi Harry Mao, Lei Luo, and Zhihong Mao. Sampling through the lens of sequential decision making. *arXiv preprint arXiv:2208.08056*, 2022c.

Sarah E Graham, Shoa L Clarke, Kuan-Han H Wu, Stavroula Kanoni, Greg JM Zajac, Shweta Ramdas, Ida Surakka, Ioanna Ntalla, Sailaja Vedantam, Thomas W Winkler, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*, 600(7890): 675–679, 2021.

Negar Hassanpour and Russell Greiner. Variational auto-encoder architectures that excel at causal inference. *arXiv preprint arXiv:2111.06486*, 2021.

Irina Higgins, Loïc Matthey, and et al. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. *arXiv preprint arXiv:1809.02383*, 2018.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *NeurIPS*, 2008.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *SIGKDD*, 2018.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization

bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *arXiv:1406.5298*, 2014.

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *PNAS*, 111(9):3354–3359, 2014.

Yann LeCun and Corinna Cortes. Mnist handwritten database. 2010.

Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment. *arXiv preprint arXiv:2012.02813*, 2020.

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 *ICML*, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *ICML*, 2020.

David A Lynch and Mustafa L Al-Qaisi. Quantitative ct in copd. *Journal of thoracic imaging*, 28(5):284, 2013.

Gianluca Maguolo and Loris Nanni. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *Information Fusion*, 76, 2021.

Haiyi Mao, Matthew J. Broerman, and Panayiotis V. Benos. Interpretable factors in scrna-seq data with disentangled generative models. In *BIBE*, 2020.

Haiyi Mao, Minxue Jia, Marissa Di, Kun Zhang, and Panayiotis V Benos. Towards hierarchical causal representation learning for nonstationary multi-omics data. *bioRxiv*, 2022a.

Haiyi Mao, Minxue Jia, Jason Xiaotian Dou, Haotian Zhang, and Panayiotis V Benos. Coem: Cross-modal embedding for metacell identification. *arXiv preprint arXiv:2207.07734*, 2022b.

Michael Mathieu, Junbo Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. Disentangling factors of variation in deep representations using adversarial training. *arXiv preprint arXiv:1611.03383*, 2016.

Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Vineet K Raghu, Allen Poon, and Panayiotis V Benos. Evaluation of causal structure

learning methods on mixed data types. In *SIGKDD Workshop on Causal Discovery*. PMLR, 2018.

Liqa A Rousan, Eyhab Elobeid, Musaab Karrar, and Yousef Khader. Chest x-ray findings and temporal lung changes in patients with covid-19 pneumonia. *BMC Pulmonary Medicine*, 20(1):1–9, 2020.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7(10), 2006.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, 2016.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv preprints*, pages arXiv–2006, 2020.

Victor Veitch, Yixin Wang, and David M Blei. Using embeddings to correct for unobserved confounding in networks. *arXiv preprint arXiv:1902.04114*, 2019.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *UAI*, 2020.

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

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523):1228–1242, 2018.

Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: disentangled representation learning via neural structural causal models. In *CVPR*, 2021.

Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021.

Kun Zhang and Aapo Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *ECMLPKDD*, 2009.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *NeurIPS*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *AISTATS*, 2020.

Appendix A. Related Work

In this section, we introduce a flurry of recent research on the intersection of representation learning and causality;we highlight the differences between existing literature and our work.

Disentangled & Causal Representation Learning. Traditional approaches to disentangled learning usually enforce the independence of different dimensions of the representations (Higgins et al., 2017; Kim and Mnih, 2018; Chen et al., 2016, 2018; Mao et al., 2020). However, this inductive bias of statistical independence is insufficient for disentanglement due to its non-identifiability (Locatello et al., 2019). Recently there are advances incorporating extra information besides images to learn identifiable representations (Wang and Jordan, 2021; Khemakhem et al., 2020; Yang et al., 2021; Locatello et al., 2020; Bouchacourt et al., 2018; Hosoya, 2018; Shen et al., 2020; Shu et al., 2019; Träuble et al., 2020; Mao et al., 2022a, 2020). iVAE (Khemakhem et al., 2020) is one of the prominent identifiable representation learning methods which utilizes the weakly-supervised learning by imposing an exponential conditional prior with auxiliary variables to guarantee identifiability. Moreover, representation learning has tremendously improved the estimation of causal inference with high-dimensional covariates and treatments (Johansson et al., 2020; Veitch et al., 2019, 2020). These works focus on learning representations to help causal inference with high-dimensional covariates. By contrast, in our work, we attempt to learn causal representations and structures jointly.

Causal Directed Acyclic Graphs (DAGs). DAG is a natural approach to represent causal structures, where directed edges represent direct cause-effect relationships. With observational data, we generally can only identify the true DAG under the *Causal Markov* assumption and *Faithfulness* assumption (Spirtes and Zhang, 2016). Formally, consider a causal DAG $\mathbf{G} = (V, E)$, where $V := \{v_1, \ldots, v_d\}$ is a set of d nodes and E is a set of directed edges. Let $(i, j) \in E$ iff there is an edge from node i to node j. For all the nodes $v_i \in V$, the joint probability V is $p(V) = \prod_i^d p(v_i|Pa(v_i))$, where $Pa(v_i)$ are the parents of v_i in $\mathbf{G}$. The functional dependence of a random variable on its parents can be described by a structural equation model (SEM) (Pearl, 2009; Peters et al., 2017; Spirtes and Zhang, 2016).

Then the corresponding SEM of $\mathbf{G}$ implies that each variable v_j can be written as a function f_j of its parents in $\mathbf{G}$, and an independent noise ϵ_j, $v_j = f_j(Pa(v_j), \epsilon_j)$.

Score-based DAG learning methods aim to find the causal DAG by optimizing a score function that involves optimization under the combinatorial acyclicity constraint. Recently, NOTEARS (Zheng et al., 2018, 2020) reformulates the score-based DAG learning problem as a continuous constrained optimization problem. The key idea is an algebraic characterization of acyclicity, which is used to minimize the least square objective, while enforcing acyclicity. Thereby the DAG is encoded as a weighted adjacent matrix $W \in \mathbb{R}^{d\times d}$. It shows that $\mathrm{tr}(e^{W\circ W}) - d = 0$

holds if and only if W is a DAG, where $\circ$ is the Hadamard product.

Intervention & ***Counterfactuals*** . An intervention on variable $\mathbf{x}$ is an operation in a system that changes only the target variable $\mathbf{x}$ and leaves other variables in the causal DAG unchanged. We denote the intervention as $do(\mathbf{x} = x)$ which sets $\mathbf{x}$ to constant x; $p(do(\mathbf{x} = x))$ denotes the resulting distribution of the intervention. Counterfactuals are defined as "what would the value of a variable be if we intervene on some variables in the causal models." Here we focus on the counterfactual images $\mathbf{x}$ obtained when we intervene on the structured variables u. We denote $\mathbf{x}(u = u')$ as the counterfactual images if we force u to take the value u'. For instance, u may represent patient COVID-19 infection status (i.e., $u = \{0, 1\}$). $\mathbf{x}(u = 0)$ is the counterfactual chest X-Ray image if the patient does not have COVID-19 given the patient does have COVID-19, accordingly, $\mathbf{x}(u = 1)$ is the counterfactual image if the patient does have COVID-19 given the patient does not have COVID-19. Then *counterfactual variance* describes the phenomenon that if u causes some representations on images $\mathbf{x}$, then $\mathbf{x}(u = 0) \neq \mathbf{x}(u = 1)$. Conversely *counterfactual invariance* delineates that if u does not cause representations on $\mathbf{x}$, then $\mathbf{x}(u = 1) = \mathbf{x}(u = 0)$.

Appendix B. Intervention Visualization

In this section, we demonstrate that the causal effects can be visualized by our model. For example, take causal case (1) $u_i \rightarrow z_j \dashrightarrow \mathbf{x}$. By Assumption 1, there exists $f_j(u_i) + e_j = z_j$, at the same time the corresponding generative function $\mathbf{g}(\mathbf{z})$ is learned. Thus we can generate the new $\tilde{\mathbf{x}}$ by setting $u_i = u_i'$. Given the assumptions of no latent confounders and mutual independence, $p(\mathbf{z}|\mathbf{z_0}, do(u_i = u_i'))$ can be visualized/estimated by the generated image $\tilde{\mathbf{x}}$ due to the following,

$$p(\mathbf{x}|do(u_i = u_i')) \propto p_\theta(\mathbf{x}|\mathbf{z}) f_j(u_i = u_i'). \quad (15)$$

For the anti-causal relation (2) $u_i \leftarrow z_j \dashrightarrow \mathbf{x}$, $\mathbf{x}$ and $\mathbf{u}$ are conditionally independent given the latent variables $\mathbf{z}$. So the the intervention of u_i will not influence on $\mathbf{x}$. Formally, the intervention on u_i can be formulated as,

$$p(\mathbf{x}|do(u_i = u_i')) = p_\theta(\mathbf{x}|u_i). \quad (16)$$

Appendix C. Theoretical Analysis

C.1. Factorized Generating Function

Explicitly, according to Figure 1(a) prior $p_\xi(\mathbf{z}|\mathbf{u})$ and posterior $q_\phi(\mathbf{z}|\mathbf{z_0}, \mathbf{u})$ can be factorized as follows:

$$\begin{aligned} p_\xi(\mathbf{z}|\mathbf{u}) &= \prod_{j=1}^{m} p_\xi^{(j)}(z_j|u_i), q_\phi(\mathbf{z}|\mathbf{z_0}, \mathbf{u}) \\ &= \prod_{j=1}^{m} q_\phi^{(j)}(z_j|\mathbf{z_0}, u_i), \end{aligned} \quad (17)$$

where u_i is the structured variable that is causally related to latent variable z_j. $p_\xi^{(j)}, q_\phi^{(j)}$ are the corresponding j-th components of p_ξ, q_ϕ respectively.

In practice, we use Gaussian distribution to model the priors,

$$p_{\boldsymbol{T},\boldsymbol{\eta}}^{(j)}(z_j|u_i) \sim \mathcal{N}(\mu_j(u_i), \sigma_j(u_i)). \quad (18)$$

Thus, given Eq. (7) & (17), the ELBO of causal representation learning module is

$$\begin{aligned} \mathcal{L}_2 &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{z_0},\mathbf{u})} \log(\frac{p_\xi(\mathbf{x}|\mathbf{z}) \prod_{j=1}^{m} p_\xi^{(j)}(z_j|u_i)}{\prod_{j=1}^{m} q_\phi^{(j)}(z_j|\mathbf{z_0}, u_i)}) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{z_0},\mathbf{u})} \log[p_\xi(\mathbf{x}|\mathbf{z}) \\ &- \sum_{j=1}^{m} D_{KL}(q_\phi^{(j)}(z_j|\mathbf{z_0}, u_i)||p_\xi^{(j)}(z_j|u_i))]. \end{aligned} \quad (19)$$

C.2. Theoretical Results

In this section, we first define the identifiability, then introduce the theorem that guarantees the causal representations identifiable

under Assumptions 1&3. In the end, we show the two-module representation learning component can asymptotically recover the identifiable causal representations.

Definition 2 ($\sim$ identifiable) *Let $\sim$ be an equivalent relation on a parameter space Θ, p_θ is identifiable up to $\sim$ that (Khemakhem et al., 2020)*

$$p_\theta(\mathbf{x}) = p_{\tilde{\theta}}(\mathbf{x}) \Rightarrow \theta \sim \tilde{\theta}.$$

Definition 3 ($\sim_B$-identifiable) *Let $\sim_B$ be the equivalence relation of the model parameter space Θ of $\{\mathbf{g}, \mathbf{T}, \boldsymbol{\eta}\}$, (Khemakhem et al., 2020)*

$$\begin{aligned} &\mathbf{g}, \mathbf{T}, \boldsymbol{\eta} \ \sim \tilde{\mathbf{g}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\eta}} \\ &\Leftrightarrow \exists B \text{ and } \mathbf{c},\ \mathbf{T}(\mathbf{g}^{-1}(\mathbf{x})) = B\tilde{\mathbf{T}}(\tilde{\mathbf{g}}^{-1}(\mathbf{x})) + \mathbf{c}, \end{aligned} \tag{20}$$

where B is the invertible permutation matrix, $\mathbf{c}$ is a vector.

we introduce following theorems to ensure $\mathbf{z}$ $\sim_B$-identifiable.

Theorem 4 (identifiability guarantee) *With Assumption 1 & 3, and a injective function $\mathbf{g}$, if $\boldsymbol{T}_{i,j}$ are differentiable almost every where, and $\boldsymbol{T}_{i,j} (1 \leq j \leq k)$ are linearly independent on any subset of $\mathbf{x}$ of measure greater than zero, then the parameters $(\mathbf{g}, \mathbf{T}, \boldsymbol{\eta})$ $\sim_B$identifiable.*

The proof of Theorem 1 is very similar to the main identifiability proof in (Khemakhem et al., 2020).

Appendix D. Algorithm Details

D.1. Main Notations

See Table E.1.

D.2. Model Architecture

See Figure 7.

D.3. Implementation

In this part we show the implementation details of our model.

- *Latent Embedding*: We use the convolutional neural network to implement the encoder and decoder. For the MNIST dataset, the encoder includes "32/32/64 /64/256/256" channels convolutional 2 times down-sampling layers, the decoder includes "256/256/64/64/32/32/32" channels with 2 times up-sampling layers. For the synthetic dataset, the encoder includes "16/16/32/32/64/64" channels convolutional 2 times down-sampling layers, the decoder includes "64/64/32/32/16/16" channels with 2 times up-sampling layers. For the X-Ray dataset, the encoder includes "32/32/64/64/256/256/512/512" channels convolutional 2 times down-sampling layers, the decoder includes "512/512/256/256/64/64/32/32/32" channels with 2 times up sampling layers. There are ReLU layers between each convolutional layers. We use the cross-entropy loss for $\log p_\phi(\mathbf{x}|\mathbf{z_0})$.

- *Causal Representation Learning*: We use MLP to implement the $p_\phi(\mathbf{z}|\mathbf{z_0}, \mathbf{u})$. MLP has 3 hidden fully connected layers with 50 neurons. The $p_\xi(\mathbf{x}|\mathbf{z})$ is implemented by a convolutional neural network, where includes "256/256/64/64/32/32/32" channels with 2 times up-sampling layers. The same as the latent embedding module, we adopt the cross-entropy loss for $\log p_\phi(\mathbf{x}|\mathbf{z})$.

- *Causal DAG*: The causal DAG learning component is implemented MLPs. Each MLP has 3 hidden layers and each layer has dimension 10. More details about

NOTEAR can be found in (Zheng et al., 2020).[3]

- *CR*: The *CR*'s classifier D component is implemented by MLPs. The MLP have 3 hidden layers and each layer has a dimension 10 and the Cross-Entropy for loss function.

Appendix E. Experiments Details

E.1. Synthesizing Data Pipeline

In order to have round and fair experiments, we designed the following data simulating pipelines. **(1)** The undirected graphs start from creating arbitrary numbers of structured variables and causal representations. Then 1-on-1 node pairs between structured variables $\mathbf{u}$ and representations $\mathbf{z}$ are randomly picked with undirected edges. Then we set a random direction for every edge. **(2)** We use non-linear causal functions to synthesize the data given the DAG structure. Specifically, for every node v_i in the graph in topological order, we use $v_i = f_i(Pa(v_i)) + \epsilon_i$, and $\epsilon_i \sim \mathcal{N}(0, 1)$ to simulate data. Similar to the settings in previous work (Zheng et al., 2020), MLPs are adopted as nonlinear functions f. **(3)** Images are generated from Gaussian Process with up-sampling functions. Specifically, we use Gaussian process to do a non-linear transformation on $\mathbf{z}$ with noises; then we formulate a random matrix with dimension 5×5 and fill $\mathbf{z}$ in at random entries; last we use interpolations to upsample the matrix dimension as $\mathbf{x}$.

E.2. Synthetic Experiments Setup

With the data synthesizing pipeline mentioned in , we simulate three datasets. For the first dataset, we simulated 10 variables where the first 5 are latent representations $\mathbf{z}$, last 5 variables are structured variables $\mathbf{u}$. Next, we randomly select $[3, 5]$ structured and latent variables pairs to connect with edges. Then we randomly set the edge directions. The simulated adjacent matrix can be found in the first row of Figure 3. The second dataset consists of more variables. For simplicity, all the variables of the simulated DAG have at most one edge. We denote a simulated DAG as n/m; there are n edges and m nodes therein. Then under our experiments settings, there are n structured variables and $m - n$ latent representations that are causally related. In total, we simulate following graphs $[1/10, 4/10, 6/12, 8/16, 10/20, 12/24, 20/40]$.

Specifically, we simulate the following seven graphs: **(1)** 1 structured variable, 9 latent variables, 1 edge between them; **(2)** 4 structured variables, 6 latent variables, 4 edges between them; **(3)** 6 structured variables, 6 latent variables, 6 edges between them; **(4)** 8 structured variables, 8 latent variables, 8 edges between them; **(5)** 10 structured variables, 10 latent variables, 10 edges between them; **(6)** 12 structured variables, 12 latent variables, 12 edges between them; **(7)** 20 structured variables, 20 latent variables, 20 edges between them. Then corresponding images are generated. The third dataset includes $[3/10, 5/10, 8/16, 10/20]$ graphs with the same settings as the second dataset.

E.3. Vertically shift of MNIST

As in Experiment B, the intervention on the causal cases 1 can be visualized in our framework. We demonstrate both causal relations discovery and interventions visualization in MNIST dataset (LeCun and Cortes, 2010). The same as rotation settings, we retrieve 10 different digits images from the dataset. We simulate a structured variable v, then shift digits images with $f(v)$ value.

3. https://github.com/aldro61/dag_with_no_tears.

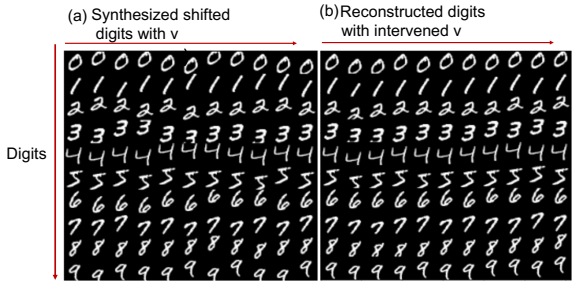

Figure 6: Intervention visualization of vertically shifting images. (a) Simulated vertically shifting images by a function of a variable $v \sim Uniform(-5, 5)$; (b). We first apply *CMCL* to discover the causal relation $v \to z_3$, where z_3 encoding the vertically shifting. Then we intervene v from from -5 to 5 and re-generate images by $p(\mathbf{x}|do(v))$. The reconstructed images are shown in (b). As expected, the reconstructed images are vertically shifted correspondingly with intervened v values. In the two figures, rows are different digits, columns are images intervened with different v values.

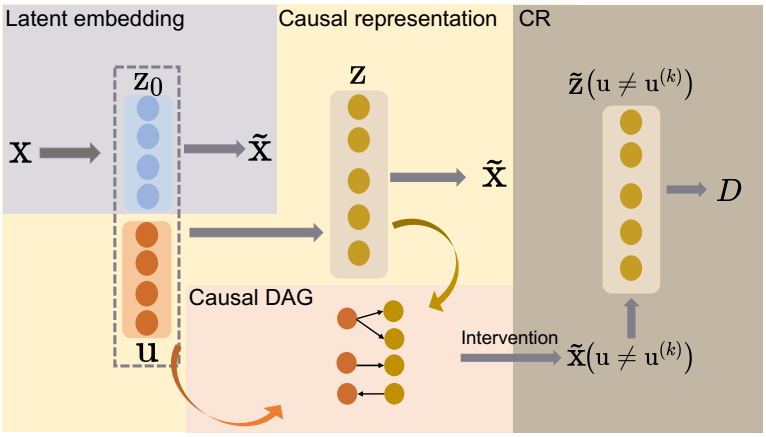

Figure 7: The architecture of *CMCL*. First, we learn the latent embeddings $\mathbf{z_0}$ of images $\mathbf{x}$. Second the causal representations are estimated by $\mathbf{z_0}$ and $\mathbf{u}$. Then we learn causal DAG with $\mathbf{z}$ and $\mathbf{u}$. Lastly we use *CR* to enforce counterfactual variance/invariance. Specifically, we use causal DAG to generate counterfactual images $\tilde{\mathbf{x}}(\mathbf{u} \neq \mathbf{u}^{(k)})$; next the counterfactual representations $\tilde{\mathbf{z}}(\mathbf{u} \neq \mathbf{u}^{(k)})$ are learned by the latent embedding and causal representation modules. Then the classifier D to distinguish the causal/anti-causal counterfactual representations. Note all these components are trained jointly.

where $f(\cdot)$ is a non-linear function according to Assumption 2. We define v and $f(\cdot)$ as: $v \sim Uniform(-5, 5)$, $f(v) \sim Norm(v, 1) + Norm(0, 1)$. Here the structured variables are $\mathbf{u} = [v, digits]$, $\mathbf{z}$ has dimension 10. In total, we simulate 1000 samples of images with corresponding v. After training the model, we find out latent variable z_3 is causally related to v as $v \to z_3$ where z_3 encodes the shifting operation. The same as the rotation experiments, we intervene v by sampling 10 different values from $v \in [-5, 5]$. Next,

Table 2: Ablation Study

CM	CR	SHD	TPR	FPR	FDR
		$12.32_{\pm 0.6}$	$0.32_{\pm 0.08}$	$0.028_{\pm 0.004}$	$0.56_{\pm 0.08}$
✓		$7.76_{\pm 0.7}$	$0.53_{\pm 0.06}$	$0.013_{\pm 0.006}$	$0.35_{\pm 0.05}$
	✓	$11.75_{\pm 0.5}$	$0.36_{\pm 0.07}$	$0.022_{\pm 0.005}$	$0.51_{\pm 0.05}$
✓	✓	$\mathbf{4.81_{\pm 0.3}}$	$\mathbf{0.77_{\pm 0.05}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.00_{\pm 0.00}}$

Table 3: Main notations and description

Notation	Type	Description
$\mathbf{x}$	Input	Images variables
$\mathbf{u}$	Input	Structured variables
$\mathbf{z_0}$	Learnable	Latent embedding of x
$\mathbf{z}$	Learnable	Causal representation, estimation of $\mathbf{z}$
$\mathbf{g}$	Learnable	Generating functions
$\mathbf{f}$	Learnable	Causal functions
$\mathbf{x}(u = u^{(0)})$	Learnable	Counterfactual images had $u = u^{(0)}$
$\mathbf{z}(u = u^{(0)})$	Learnable	Counterfactual $\mathbf{z}$ had $u = u^{(0)}$
W	Learnable	Adjacent matrix of DAG
ϕ	Learnable	Variational parameters
ψ	Learnable	Priors in latent embedding module
ξ	Learnable	Priors in causal representation module
$\mathbf{T}, \boldsymbol{\eta}$	Learnable	Parameters of exponential distribution
Ω^+	Learnable	Set of $u_i \in \mathbf{u}$ are parents of $z_j \in \mathbf{z}$
Ω^-	Learnable	Set of $u_i \in \mathbf{u}$ are children of $z_j \in \mathbf{z}$

we use learned relations between v and z_3 to infer the new $\mathbf{z}$ and generate new images. The generated images are shown in Fig. 6. From the results, we correctly reconstruct intervened images. As expected, the reconstructed images are vertically shifted correspondingly with intervened v values.

E.4. Ablation Study

We conducted ablation study on the synthesis dataset in Experiment section with 40 variables and 20 edges. The results of SHD (Structure Hamming Distance), TPR (True Positive Rate), FPR (False Positive Rate), and FDR (False Discover Rate). We repeated this experiments for 5 times and reported the mean and standard deviation in following table. We use CM to denote causal representation learning module, CR to denote the contrastive regularizer. From the ablation study results Table 2, both Causal learning module and contrastive regularizer help improve the DAG learning. The causal learning module can contribute to most performance improvement. Note that for model without CM module, we directly flatten the $\mathbf{z}_0$ as $\mathbf{z}$.

E.5. Procedures

Here we provide the pseudo code of our *CMCL* algorithm in Algorithm 1.

Algorithm 1: Cross-Modal Variational Causal representation and structure Learning

Input images: $\mathbf{x}$, structured variables: $\mathbf{u}$.
Outputp: generating function $\mathbf{g}$, causal representations:$\mathbf{z}$, DAG adjacent matrix:W, causal functions $\mathbf{f}$.

1. Initialization ϕ, W, A
2. **While** Training
   (a) Training Latent Embedding
      i. $\mathbf{z_0} \sim q_\phi(\mathbf{z_0}|\mathbf{x})$
      ii. *Minimize* $\mathcal{L}_1$ (6)
      iii. *Update*(ϕ)
   (b) Causal Representation Learning
      i. $\mathbf{z_0} \sim q_\phi(\mathbf{z_0}|\mathbf{x})$
      ii. *Minimize* $\mathcal{L}_1$ (6)
      iii. *Update*(ϕ)
   (c) Training DAG
      i. *train_DAG*($\mathbf{z}, \mathbf{u}$) (10)
      ii. *update*(W, A, ϕ)
   (d) Training *CR*
      i. *Sampling* $u_i \in \Omega_+$ *and* $u_j \in \Omega_-$
      ii. *Intervene* u_i, u_j;
      iii. *Generate* $\tilde{\mathbf{x}}^{(k)}(u_i \neq u_i^{(k)}), \tilde{\mathbf{x}}^{(k)}(u_j \neq u_j^{(k)})$.
      iv. *Compute* $\tilde{\mathbf{z}}^{(k)}(u_i \neq u_i^{(k)}), \tilde{\mathbf{z}}^{(k)}(u_j \neq u_j^{(k)})$.
      v. *Train* $\mathbf{D}$ (13)
3. **End While**