SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge

Aneesh Komanduri[1], Yongkai Wu[2], Wen Huang[1], Feng Chen[3], and Xintao Wu[1]
[1]*University of Arkansas*, Fayetteville, Arkansas, USA
[2]*Clemson University*, Clemson, South Carolina, USA
[3]*University of Texas at Dallas*, Richardson, Texas, USA
akomandu@uark.edu, yongkaw@clemson.edu, wenhuang@uark.edu,
feng.chen@utdallas.edu, xintaowu@uark.edu

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10021114

***Abstract*—The goal of causal representation learning is to map low-level observations to high-level causal concepts to learn interpretable and robust representations for various downstream tasks. Latent variable models such as the variational autoencoder (VAE) are frequently leveraged to learn disentangled representations. However, there are often complex non-linear causal relationships underlying the observed data that cannot be captured through disentangled representations or linear dependence assumptions. Further, an independent conditional prior assumption can make learning causal dependencies in the latent space more challenging. We propose a framework, coined SCM-VAE, which uses apriori causal knowledge, a structural causal prior, and a non-linear additive noise structural causal model (SCM) to learn independent causal mechanisms and identifiable causal representations. We conduct theoretical analysis and perform experiments on synthetic and real-world datasets to show the improved quality of learned causal representations and robustness under interventions.**

***Index Terms*—causality, variational autoencoders, representation learning**

I. Introduction

Causality [1] has profoundly influenced how we think about modern AI problems and it has been argued that causality is crucial for reasoning about the world [2]. Recently, there has been a growing interest at the intersection of causality and machine learning to learn causal models of the world. Causality has proven to be quite useful in a variety of domains including algorithmic fairness and out-of-distribution generalization. As such, causal modeling can often be used to learn trustworthy, robust, and explainable machine learning models. Representation learning [3] has seen significant development over the past decade. However, there has been little work in bringing causality to representation learning to learn meaningful abstract representations from data. Recent work has focused on learning high-level causal representations [4]–[6] to explain low-level observational features. Learning such representations can be useful for tasks involving scheduling, planning, and robustness to distribution shifts [7]. Disentangled representation learning [8]–[10], which aims to encode data into independent factors of variation, has been often linked with causal representation learning, where the factors of variation may be causally dependent. Such representations could be consistent with a causal generative process in that an independent mechanism generates each factor. However, learning causal representations remains to be a challenging problem since arbitrary and complex neural networks are unable to predict the effects of interventions given only observational data. Thus, introducing a new form of inductive bias that encodes causal structure information is necessary to bridge the gap between conventional machine learning and causal inference [11]. Therefore, learning causal representations with interventional and counterfactual generation capabilities is vital to achieving robustness for downstream tasks.

Many existing works have explored disentangled representation learning as a means to learn causal representations. Learning identifiable disentangled representations requires a weak supervision signal [12] and often assumes the causal structure or a super-graph underlying the causal variables is given apriori [13]. A recent endeavor [5] proposed a VAE-based causal disentangled representation learning framework (CausalVAE) that tries to learn causal representations while automatically discovering the causal structure. However, this method has several limitations. For one, the linear SCM assumption is unrealistic and fails to capture complex causal dependencies. Second, CausalVAE leverages the Mask Layer developed in [14] that, unfortunately, can only discover a super-graph of the true causal graph. Further, the joint constraint-based approach with causal discovery is sensitive to the method of initialization of the adjacency matrix, making it an unstable approach to learning the underlying causal structure and causal representation.

Therefore, we assume a matrix of topologically ordered causal variables is available apriori, e.g., either from domain knowledge, extracted from labeled data, or some causal discovery method. We follow this assumption for two reasons. Firstly, apriori causal knowledge is a valid assumption since the user often specifies semantically meaningful variables and prior knowledge helps guide meaningful representations. Secondly, joint optimization approaches that combine causal discovery and causal representation learning using the acyclicity constraint often have unidentifiability issues, can be difficult to tune in practice, and at best recover only the super-graph [14]. Further, several works in the literature, such as [13] and [15], also assume causal orderings or structure is given apriori to learn causal representations. We take a similar approach and

focus on learning meaningful causal representations from prior knowledge about the structure between causal variables.

We develop a framework, coined SCM-VAE, in which we assume a non-linear structural causal model. We propose a structural causal prior, a key component of our approach, which regularizes the posterior and enforces the causal structure among dimensions of the latent representation consistent with the causal graph. This addresses another concern of CausalVAE, where the conditional factorized prior simply assumes mutual independence among factors. Further, we rigorously show that the causal representation and causally related supervision labels must have the same causal structure, assuming they have a one-to-one correspondence. We design the generative model such that the latent dimensions are generated by independent causal mechanisms that are ensured through the causally factorized prior. Instead of learning causal representations through constraints on the loss function, we target learning the causal representation directly through the structural causal prior and binary adjacency matrix. We show that our framework is compatible with performing interventions and generating counterfactual instances.

Our contributions are as follows. (1) We develop SCM-VAE, a framework for causal representation learning under apriori causal structure knowledge, assuming a non-linear additive noise structural causal model. (2) Further, we propose a causally factorized structural causal prior based on the known topological orderings of the causal graph to enforce causal representations. (3) We conduct theoretical identifiability analysis and perform experiments on synthetic and real-world datasets to demonstrate the improved quality of learned causal representations under apriori causal knowledge and robustness under interventions.

II. Related Work

Disentangled Representation Learning. Disentangled representation learning methods focus on learning mutually independent factors of variation. Latent variable models such as variational autoencoders (VAE) are a common framework used to learn disentangled representations. The goal of disentangled representation learning is to encode data into a latent space and approximate the true posterior distribution $p(\mathbf{z}|\mathbf{x})$ of the latent space by using a variational distribution $q(\mathbf{z}|\mathbf{x})$ and ensure that the dimensions of the latent variable are mutually independent through some independent prior used to regularize the posterior. There have been several modifications of the original VAE objective proposed for disentangled representation learning such as ConditionalVAE [16] and β-VAE [17]. However, some methods, such as β-VAE, are unsupervised and do not have identifiability guarantees. A critical component of disentangled representations is to show that they are identifiable. That is, the model can approximate the true parameters up to some trivial transformation. [8] developed a theory of identifiability for variational autoencoders using a prior conditioned on auxiliary labels, which builds on the fundamental principles of non-linear Independent Component Analysis (ICA) [18].

Causal Disentangled Representation Learning. A desirable property of causal representations is the independence of causal mechanisms. Recent work has shown that learning disentangled representations can help achieve this property in causal representations. [19] developed the idea of a disentangled causal process and learning disentangled latent representations that are all high-level direct causes of the low-level data. [20] proposed a causal implicit generative model called CausalGAN that considers the problem of causal controllable generation based on a given dependency structure between labels to allow for controlled observational and interventional generative capabilities. However, this work focuses only on controlled generation rather than representation learning. [5] proposed a latent variable model called CausalVAE, which uses the formulations of DAG-GNN [21] and Masked Causal Structure Learning [14] to learn a causal graph as a part of a pretraining procedure and learn causal disentangled representations using a variational autoencoder (VAE) weakly supervised by a conditional prior. CausalVAE also supports interventions by including an SCM masking layer to propagate the effects of parent nodes to children nodes. Later, [13] generalized the formulation of CausalVAE to bidirectional generative models and proposed an SCM causal prior to achieve identifiable disentangled representations given causal orderings apriori. [22] introduced the concept of latent causal models (LCMs) and proposed to learn causal representations and causal structure by introducing interventional data into the training process as weak supervision.

Causal Discovery. The problem of structure learning of DAGs has proven to be quite a difficult task, primarily due to combinatorial optimization methods that have high complexity. There has been much attention on finding gradient-based solutions to DAG structure learning. Continuous optimization methods have recently been proposed to solve structure learning problems using gradients. NOTEARS [23] formulates the problem of learning the structure of DAGs as optimizing a continuous program with respect to an acyclicity constraint enforced on the graph to estimate a weighted adjacency matrix and ensure the graph is indeed a DAG. Second-order methods, such as the augmented Lagrangian, are used to optimize the objective and learn the true structure of the DAG. DAG-GNN [21] is an extension of this formulation to a non-linear setting and uses a deep latent variable model and graph neural networks to model the structure of the DAG. This work further proposes an adaptation of the acyclicity constraint and augmented Lagrangian to a deep learning setting.

III. Preliminaries

A. Structural Causal Model

A structural causal model (SCM) [24] is formally defined by a triple $\mathcal{M} = \langle \mathbf{Z}, \mathbf{N}, \mathbf{F} \rangle$, where $\mathbf{Z}$ is the set of n endogenous variables, $\mathbf{N}$ is a set of n exogenous independent noise variables, and $\mathbf{F}$ is a collection of n structural equations of the form:

$$Z_j := f_j(\mathbf{Pa}_j, N_j), j = 1, \ldots, n \tag{1}$$

where $\mathbf{Pa}_j \subseteq Z \setminus \{Z_j\}$ are called parents or direct causes of Z_j and the exogenous noise N_j ensures to represent a general conditional distribution $P(Z_j|\mathbf{Pa}_j)$. An SCM where the exogenous noise variables are jointly independent (no hidden confounders) is known as a Markovian model, which is the setting we assume for the purposes of this work. We depict an SCM $\mathcal{M}$ graphically by a directed acyclic graph (DAG) known as a causal graph $\mathcal{G} = (V, E)$, where each node in V is an endogenous variable and each directed edge in E represents a direct causal relationship between parent and child as defined by the structural equations. In this work, we assume the additive noise model, $Z_j := f_j(\mathbf{Pa}_j) + N_j$ for $j = 1, \ldots, n$, where f_j is a deterministic non-linear function and N_j's are mutually independent noise variables with strictly positive densities.

As formulated in [4], causal representation learning aims to learn interpretable causal variables from raw data and the connection from $(z_1, \cdots, z_n)$ to raw observation $\mathbf{x}$ can generally be expressed as $\mathbf{x} = G(z_1, \cdots, z_n)$ where G is a non-linear function. The Independent Causal Mechanisms (ICM) principle implies

$$p(z_1, \cdots, z_n) = \prod_{j=1}^{n} p(z_j|\mathbf{Pa}_j) \tag{2}$$

where mechanisms $z_j := f_j(\mathbf{Pa}_j, N_j), j = 1, \cdots, n$ model the causal relationships among z_j.

B. Variational Autoencoders

Let $\mathbf{x} \in \mathbb{R}^d$ be an observed data variable, $\mathbf{z} \in \mathbb{R}^n$ be the latent causal representation. We have the deep latent variable model as:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z}) \tag{3}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a vector of parameters, $p_{\boldsymbol{\theta}}(\mathbf{z})$ is the prior distribution over the latent variables. $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is parameterized with a neural network called the decoder. Suppose the observed dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\}$ is generated from $p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}^*}(\mathbf{z})$ where $\boldsymbol{\theta}^*$ are the true but unknown parameters. The VAE [25] learns a full generative model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z})$ and an inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ that approximates its posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. The VAE model can efficiently optimize the parameters $\boldsymbol{\theta}$ to have $p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z})$. However, it cannot achieve the approximation of the joint distribution, $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z})$.

In [8], the authors define a family of identifiable deep latent variable models, called iVAE, that can learn the true joint distribution. The iVAE assumes a conditionally factorized prior distribution over the latent variables $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ where $\mathbf{u} \in \mathbb{R}^n$ is an additionally observed supervision variable such as class labels. Let $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{T}, \lambda)$ be the parameters of the conditional generative model:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_{\mathbf{f}}(\mathbf{x}|\mathbf{z})p_{\mathbf{T},\lambda}(\mathbf{z}|\mathbf{u}) \tag{4}$$

The first term is defined as $p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) = p_{\epsilon}(\mathbf{x} - \mathbf{f}(\mathbf{z}))$. In other words, the value of $\mathbf{x}$ can be decomposed as $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \epsilon$ where ϵ is an independent noise variable with PDF $p_{\epsilon(\epsilon)}$. The function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^d$ is injective and can be any non-linear function which in practice is approximated by neural networks.

Assuming the observed dataset $\mathcal{D} = \{(\mathbf{x}^{(1)}, \mathbf{u}^{(1)}), \cdots, (\mathbf{x}^{(N)}, \mathbf{u}^{(N)})\}$ is generated according to the generative model defined in Eq. 4, the VAE can be used to learn the true generating parameters $\boldsymbol{\theta}^*$. The VAE learns a deep latent generative model and a variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ of its true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$. Denote the conditional marginal distribution of the observations as $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u})d\mathbf{z}$ and denote the empirical data distribution given $\mathcal{D}$ as $q_{\mathcal{D}}(\mathbf{x}, \mathbf{u})$. The VAE learns the vector of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ by maximizing

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) := \mathbb{E}_{q_{\mathcal{D}}}\Big[\mathbb{E}_{q_{\boldsymbol{\phi}(\mathbf{z}|\mathbf{x},\mathbf{u})}}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) \\ - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})\big]\Big]\end{aligned} \tag{5}$$

IV. SCM-VAE

We start with an apriori causal graph we assume is given from domain knowledge and propose our framework SCM-VAE to achieve robust causal representations. We define an additive noise SCM to represent the causal generative process and encode the structure in a structural causal prior, which is used as weak supervision to learn the latent posterior distribution of a latent variable model. We analyze the properties of the causal prior to justify the encoding of causal representations. Further, we show the interventional capabilities of our model through a causal controllable generation mechanism.

A. Encoding Causal Structure via Additive Noise SCM

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the binary adjacency matrix of a causal graph consistent with some n-variable causal generative process. The matrix $\mathbf{A}$ is an upper triangular matrix consisting of topologically sorted causal ordering. Further, we assume that the same causal structure is encoded in the latent causal representation modeled as a general additive noise non-linear reduced form SCM:

$$\mathbf{z} = \mathbf{g}(\mathbf{A}, \mathbf{z}) + \boldsymbol{\epsilon} \tag{6}$$

where $\mathbf{z} \in \mathbb{R}^n$ is the latent causal representation of endogenous variables $\mathbf{u}$, $\boldsymbol{\epsilon}$ is independent exogenous random noise sampled from an n-variate Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{g}$ is a set of non-linear functions. Specifically, let $\mathbf{A} = [\mathbf{A}_1|\mathbf{A}_2 \cdots |\mathbf{A}_n]$ be the binary adjacency matrix associated with the true DAG $\mathcal{G}$ where A_{ji}, the j-th entry of $\mathbf{A}_i$, equals 1 if and only if z_j is a parent of z_i. In our formulation, the ANM can be rewritten in a form parameterized by $\mathbf{A}$:

$$z_i = g_i(\mathbf{A}_i \odot \mathbf{z}) + \epsilon_i \tag{7}$$

for all $i = 1, \ldots, n$, where z_i is evaluated according to an assumed causal topological ordering of n concepts, $\odot$ is the element-wise product, and g_i's are learned independent causal mechanisms via *separate* neural networks. For root nodes, $z_i = \epsilon_i$, the representation from the encoder. [14] shows under suitable conditions the above formulation results in identification of a super-graph of the true graph, from which

non-linear variable selection methods can be used to derive the parental sets and learn the causal graph. CausalVAE leverages this Mask Layer that can only discover a super graph of the true causal graph.

We aim to learn the causal representation of high-level endogenous variables using the causal structure $\mathbf{A}$ and a latent variable model. Our model aims to map each latent dimension of the causal representation $\mathbf{z}$ to the corresponding dimension of the endogenous variables $\mathbf{u}$ while encoding the causal structure among the latent dimensions. Learning the causal representation consistent with the given causal ordering allows a decoder to perform counterfactual generation of new data through strategic interventions on causal variables. In practice, the exogenous random noise $\boldsymbol{\epsilon}$ is captured by the variance of the Gaussian distribution of the sampled latent variable.

B. Learning Causal Representations with Variational Autoencoder

We develop our model using the variational autoencoder (VAE) generative framework [25] to learn robust causal representations with interventional capabilities. Let $\mathbf{u} \in \mathbb{R}^n$ be the set of observed variables of interest and each element u_i be an endogenous variable in the causal graph. Let $\mathbf{z}$ be the latent causal representation of endogenous variables that represent the high-level semantics of the data and $\boldsymbol{\epsilon}$ be the intermediate latent representation. Let $\mathbf{A}$ be the triangular matrix consisting of topologically sorted causal ordering.

We consider the following conditional generative model parameterized by $\boldsymbol{\theta} = (\mathbf{E}, \mathbf{D}, \mathbf{A}, \mathbf{T}, \boldsymbol{\lambda})$:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\epsilon}|\mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u}) p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}) \tag{8}$$

Let $\mathbf{E}$ denote the encoder and $\mathbf{D}$ denote the decoder. We assume the following decoding and encoding processes:

$$\mathbf{x} = \mathbf{D}(\mathbf{z}) + \xi, \qquad \boldsymbol{\epsilon} = \mathbf{E}(\mathbf{x}, \mathbf{u}) + \zeta \tag{9}$$

where ξ and ζ are the vectors of independent noise with probability densities p_ξ and q_ζ. In VAE, an inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ approximates the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$.

Then, the probabilistic encoder, parameterized by variational parameters $\boldsymbol{\phi}$, is as follows:

$$q_{\boldsymbol{\phi}}(\mathbf{z}, \boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u}) \equiv q(\mathbf{z}|\boldsymbol{\epsilon}) q_\zeta(\boldsymbol{\epsilon} - \mathbf{E}(\mathbf{x}, \mathbf{u})) \tag{10}$$

and the probabilistic decoder, parameterized by generative parameters $\boldsymbol{\theta}$, is as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \equiv p_\xi(\mathbf{x} - \mathbf{D}(\mathbf{z})) \tag{11}$$

Non-noisy observations $\mathbf{x} = \mathbf{D}(\mathbf{z})$ are a special case of setting $p_\xi(\xi)$ with infinitesimal variance.

The joint prior $p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})$ (the second term in Equation 8) for latent variables $\mathbf{z}$ and ϵ is defined as

$$p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) \tag{12}$$

where $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is the structural causal prior.

C. Structural Causal Prior

The structural causal prior of latent causal representations of endogenous variables is factorized as a function of the information from the given causal graph as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(z_i|u_i, \mathbf{Pa}_i(\mathbf{u})) \tag{13}$$

where

$$\begin{aligned} p_{\boldsymbol{\theta}}(z_i|u_i, \mathbf{Pa}_i(\mathbf{u})) = \mathcal{N}\big(&\boldsymbol{\lambda}_1((\mathbf{A} + \mathbf{I})_i \odot \mathbf{u}), \\ &\boldsymbol{\lambda}_2((\mathbf{A} + \mathbf{I})_i \odot \mathbf{u})\big) \end{aligned} \tag{14}$$

where $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are arbitrary functions $\mathbb{R}^d \to \mathbb{R}$ and $\mathbf{I}$ is the identity matrix. $\mathbf{Pa}_i(\mathbf{u})$ is the set of parents of u_i, the counterpart of z_i in the causal graph $\mathbf{A}$. The structural causal prior encodes information about the causal structure and is used in regularizing the posterior to enforce causal structure in representation learning. The minimization of $KL(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}))$ will enforce causal structure for the encoder.

The structural causal prior defined in Equations (13) and (14) is a special case of the conditional prior formulation proposed in [8], where each latent variable is conditioned on auxiliary labels $\mathbf{u}$. Specifically, the prior on the latent variables $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ in [8] is assumed to be conditionally factorial, where each element of $z_i \in \mathbf{z}$ has a univariate exponential family distribution given conditioning variable $\mathbf{u}$. The conditioning on $\mathbf{u}$ is through an arbitrary function $\lambda(\mathbf{u})$ that outputs the individual exponential family parameters $\lambda_{i,j}$. The PDF is given by:

$$p_{\mathbf{T},\lambda}(\mathbf{z}|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\mathbf{u})} exp\Big[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(\mathbf{u})\Big] \tag{15}$$

where Q_i is the base measure, $Z_i(\mathbf{u})$ is the normalizing constant, $\mathbf{T}_i = (T_{i,1}, \cdots, T_{i,k})$ are the sufficient statistics, and $\lambda_i(\mathbf{u}) = (\lambda_{i,1}(\mathbf{u}), \cdots, \lambda_{i,k}(\mathbf{u}))$ the corresponding parameters, and k is the fixed dimension of each sufficient statistics.

In our case, each latent variable is conditioned on its corresponding label dimension and its immediate parents. That is, we can express the prior as the Bayesian network factorization of the causal graph. We use an isotropic Gaussian distribution from the exponential family of distributions for our prior with two sufficient statistics $\mathbf{T}_i = (T_{i,1}, T_{i,2})$. Further, our structural causal prior follows a similar form to the causal (disentangled) factorization from [4].

Our formulation is also different from the prior introduced in CausalVAE [5], which simply assumes mutual independence among factors. Similar to CausalVAE, we consider the dimensions of each representation z_i to be causally related, i.e., determined by its parents, and factorize the causal prior based on the SCM structure. By encoding the graphical model in the causal prior, we encourage the learned representation to be causally related instead of independent. In addition, the learned SCM-VAE model generates high-quality interventional data when the interventional queries are performed on the

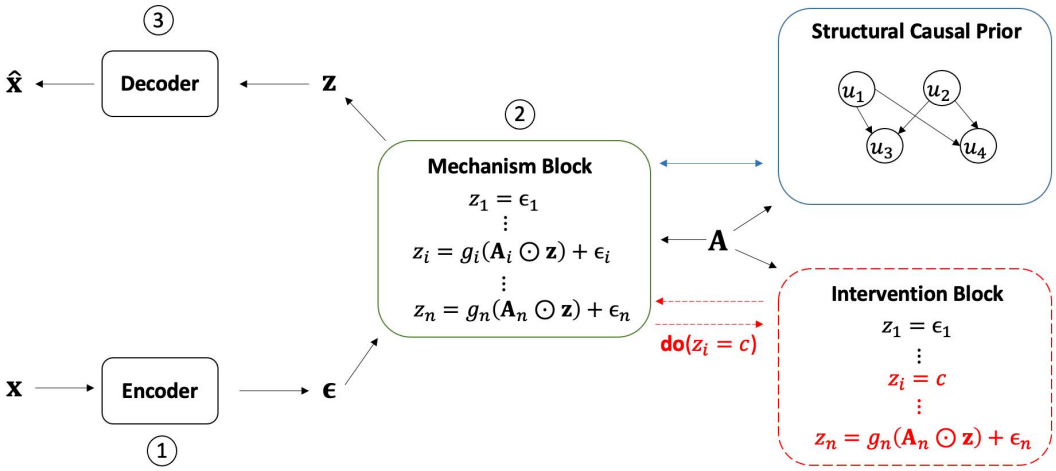

Figure 1. Our SCM-VAE Framework consists of three main components: mechanism, structural causal prior, and intervention. The input image is (1) encoded into an $(n \times k)$-dimensional noise encoding $\boldsymbol{\epsilon}$, which is then (2) transformed via the independent causal mechanisms into causal representation $\mathbf{z}$ weakly supervised by labels $\mathbf{u}$ and causal structure $\mathbf{A}$ through the structural causal prior. (3) This representation is then fed as an input to the decoder to reconstruct the image. The intervention block is only carried out during inference by intervening on a dimension of the causal representation with a constant c, recursively update representation of descendants via causal mechanisms, and generate the counterfactual instance.

representation due to the causal relationships among dimensions of the representation. In practice, u_i is location-scale normalized. The following theorem establishes a relationship between causal representation $\mathbf{z}$ and causal labels $\mathbf{u}$ and their causal factorizations as defined by our proposed prior.

Theorem 1. *If there is a bijection between representation vector* $\mathbf{z}$ *and supervision label* $\mathbf{u}$ *and we know the Bayesian network factorization of* $\mathbf{z}$ *consistent with the causal adjacency matrix* $\mathbf{A}$*, then the corresponding elements of* $\mathbf{u}$ *induce the same Bayesian network factorization under* $\mathbf{A}$*.*

The proof of Theorem 1 can be found in Appendix A. Note that although we fix the causal graph and use causal structure information in our prior, $\mathbf{A}$ remains as a parameter of our generative model since there can be constraints placed on $\mathbf{A}$ to make it a learnable parameter along with the prior. The overall architecture of our framework is shown in Figure 1.

D. Evidence Lower Bound

To learn the distribution of the latent endogenous posterior, we target learning a variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}, \boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})$ as an approximation to the true latent posterior $p_{\boldsymbol{\theta}}(\mathbf{z}, \boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})$. We learn the variational and true posterior parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively, by maximizing the following variational lower bound:

$$\begin{aligned}\mathbb{E}_{q_{\mathbf{X}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})] \geq \mathbb{E}_{q_{\mathbf{X}}}[\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\epsilon}, \mathbf{z}, \mathbf{u})] \\ - \mathcal{D}(q_{\boldsymbol{\phi}}(\mathbf{z}, \boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}))]\end{aligned} \tag{16}$$

where $q_{\mathbf{X}}(\mathbf{x}, \mathbf{u})$ is the joint distribution over the dataset $\mathbf{X}$ and supervision labels $\mathbf{u}$ and $\mathcal{D}$ is the KL divergence. Here, $\boldsymbol{\epsilon}$ and $\mathbf{z}$ have a one-to-one correspondence so that we can split the intractable joint distribution into two tractable conditional distributions as in Eq. (17). Formally, we have the following objective:

$$\begin{aligned}\max \mathbb{E}_{q_{\mathbf{X}}}[\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \mathbf{u})] \\ - \alpha\mathcal{D}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})) \\ - \beta\mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}))]\end{aligned} \tag{17}$$

E. Intervention on SCM-VAE

During inference, we are interested in the performance of our model under interventional queries. In order to achieve a good causal representation, interventions are necessary. In the Pearl Causal Hierarchy [26], we can easily achieve level 1, observational inference, by a model such as Conditional-VAE. Our model can leverage the $\mathbf{do}(\cdot)$ operator to perform interventions on dimensions of the latent variable $\mathbf{z}$ to achieve level 2: interventional inference. By intervening on a concept dimension, we effectively cut the tie to the variable's parents and replace the causal mechanism that generates the dimension with a constant value c. For instance, if we want to intervene on latent dimension i, we would perform the intervention via $\mathbf{do}(z_i = c)$. Formally, we perform interventions on the latent variable by sampling from an interventional distribution facilitated by the truncated factorization:

$$\mathbf{z} \sim P_{z_j}(\mathbf{z}|\mathbf{u}) = \prod_{i \neq j} P(z_i|u_i, \mathbf{Pa}_i(\mathbf{u}))\delta_{Z_j = z_j} \tag{18}$$

Performing the interventions, forward sampling, and evaluating the structural equations yields representations consistent with the assumed causal model. If such an intervention changes the representation consistent with the causal graph, we can generate counterfactual instances from the distribution $P(\hat{X}|\mathbf{do}(z_i = c))$, level 3, using the trained decoder. Intervention on a single concept does not influence any other non-causal concept. If we intervene on a concept of interest, only descendants of the concept in the causal graph will change

accordingly. Therefore, our model can capture independent and modular mechanisms, a desirable property of causal representations.

V. Identifiability Analysis

We derive our identifiablity analysis for SCM-VAE following the proof logic in [5] and extend the identifiability result for casual representation learning under structural casual prior. The proof of Theorem 2 can be found in Appendix B.

Definition 1. *Let $\sim$ be the equivalence relation on Θ defined as follows:*

$$(\mathbf{E}, \mathbf{D}, \mathbf{A}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\tilde{\mathbf{E}}, \tilde{\mathbf{D}}, \tilde{\mathbf{A}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}) \Leftrightarrow$$
$$\exists \mathbf{B_1}, \mathbf{B_2}, \mathbf{b_1}, \mathbf{b_2} \mid \mathbf{T}(\mathbf{E}(\mathbf{x})) = \mathbf{B_1}\tilde{\mathbf{T}}(\tilde{\mathbf{E}}(\mathbf{x})) + \mathbf{b_1}, \quad (19)$$
$$\mathbf{T}(\mathbf{D}^{-1}(\mathbf{x})) = \mathbf{B_2}\tilde{\mathbf{T}}(\tilde{\mathbf{D}}^{-1}(\mathbf{x})) + \mathbf{b_2}, \quad \forall \mathbf{x} \in \mathcal{X}$$

If $\mathbf{B_1}$ is an invertible matrix and $\mathbf{B_2}$ is an invertible diagonal matrix with diagonal elements associated to $\mathbf{u}$ and $\mathbf{A}$. The model parameter is defined to be $\sim$- identifiable.

Theorem 2. *Assume the observed data is sampled from a generative model defined according to Equations (8)-(11), with parameters $(\mathbf{E}, \mathbf{D}, \mathbf{A}, \mathbf{T}, \boldsymbol{\lambda})$. Assume the following holds:*

1) *The set $\{x \in \mathcal{X} | \phi_\xi(x) = 0\}$ has measure zero, where ϕ_ξ is the characteristic function of the density p_ξ defined in Eq. (11).*
2) *The decoder function $\mathbf{D}$ is differentiable and the Jacobian matrix of $\mathbf{D}$ is of full rank.*
3) *The sufficient statistics $T_{i,s}(z_i)$ related to Equation (14) are non-zero almost everywhere for all $1 \leq i \leq n$ and $1 \leq s \leq 2$ assuming a Gaussian distribution with sufficient statistics mean and variance.*
4) *The prior probability density is conditionally factorial and takes the form as shown in Equation (13). In Equation (22) the values $\lambda_s^i(\mathbf{u}, \mathbf{A}) \neq 0, \forall i, s$.*

Then, the parameters are $\sim$-identifiable.

VI. Experiments

In this section, we perform experiments on three datasets and compare the performance of our SCM-VAE model with the CausalVAE [5] and ConditionalVAE [16] baseline models. Note that since our setting assumes a fixed causal graph with known topological orderings, we evaluate CausalVAE with known causal structure and without any causal discovery components as denoted by the asterisk. Further, we also show the performance of our model under interventions. We run our experiments on an Ubuntu 20.04 workstation having eight NVIDIA Tesla V100-SXM2 GPUs with 32GB RAM. We run our model for 100 epochs and tune three hyperparameters: $\alpha = 0.1$ and $\beta = 1.0$ for weighing the divergence terms and $\eta = 0.001$ to scale the variance of the sampled causal representation. We assume unit variance for the structural causal prior. The encoder/decoder architecture is a 3-layer MLP using ELU activation for the Pendulum and Flow datasets and a 6-layer CNN using ReLU activation for CelebA datasets. We make our code publicly available for reproducibility.[1]

[1] https://github.com/Akomand/SCM-VAE

A. Datasets

Pendulum. We use a synthetic dataset, Pendulum, from [5], which consists of $7,000$ images (6K to train and 1K to test) generated using four continuous variables that simulate a closed pendulum system with mechanisms generating shadows as a function of light position and pendulum angle. The physical system consists of the factors *pendulum angle*, *light position*, *shadow length*, and *shadow position*, which are used as the labels for each image. We view these variables as the causal concepts of interest whose causal graph is *(pendulum angle, light position)* $\rightarrow$ *shadow length* and *(pendulum angle, light position)* $\rightarrow$ *shadow position*.

Flow. We use another synthetic dataset, Flow, from [5], which consists of $8,000$ images (6K to train and 2K to test) generated using four continuous variables that simulate a water flow system from a cup of water and a ball. The physical system consists of the factors *ball size*, *water height*, *hole*, and *water flow* which are used as the labels for each image. We view these variables as the causal concepts of interest whose causal graph is *ball size* $\rightarrow$ *water height* and *(water height, hole)* $\rightarrow$ *water flow*.

CelebA. The CelebA dataset [27] consists of $200,000$ images of celebrity faces with 40 discrete attributes with values in $\{-1, 1\}$ describing each image. We use two subsets of CelebA: CelebA-Smile and CelebA-Beard. The CelebA-Smile and CelebA-Beard consist of $20,000$ images (17K to train and 3K to test) sampled from the dataset using only the attributes *gender*, *smile*, *narrow eyes*, *mouth slightly open* and *age*, *gender*, *bald*, *beard*, respectively. The causal graph of CelebA-Smile is *gender* $\rightarrow$ *narrow eyes* and *smile* $\rightarrow$ *(narrow eyes, mouth open)*, *mouth open* $\rightarrow$ *narrow eyes*. The causal graph of CelebA-Beard is *(age, gender)* $\rightarrow$ *bald* and *(age, gender)* $\rightarrow$ *beard*.

B. Metrics

Information Coefficients [28]. Maximal Information Coefficient (MIC) is an information measure that can quantify the degree of alignment between variables. Total Information Coefficient (TIC) is a statistic that tests for independence and is the sum of the entries of the equicharacteristic matrix. We measure the alignment between the learned causal representation $\mathbf{z}$ and the ground truth causal structured labels $\mathbf{A} \odot \mathbf{u}$ using the MIC and TIC as defined in [28].

Interventional Reconstruction. To evaluate the performance of counterfactual images, we measure the average reconstruction error between the intervened image and the ground-truth intervention from the test dataset. Since we know the data generating process of the Pendulum and Flow datasets, we are able to compare the ground-truth intervention with the image reconstructed by the trained decoder.

C. Results

We evaluate our model using the MIC and TIC for representation quality and interventional reconstruction error for performance under intervention. We can learn more accurate causal representations by changing the prior to incorporating

Metrics (%)						
	SCM-VAE		CausalVAE*		ConditionalVAE	
Dataset	MIC	TIC	MIC	TIC	MIC	TIC
Pendulum	**96.2**±0.6	**89.1**±1.4	82.3±1.6	60.1±1.1	93.8±3.3	80.5±1.4
Flow	**97.9**±0.5	**90.6**±1.2	96.8±0.9	88.2±1.3	75.5±2.3	56.5±1.8
CelebA-Smile	75.1±3.8	**68.9**±3.6	67.9±3.3	63.2±3.5	**78.8**±10.9	66.1±12.1
CelebA-Beard	**94.4**±1.1	**88.9**±1.3	91.3±1.2	85.1±2.0	89.8±6.2	78.7±7.7

TABLE I: Maximal Information Coefficient (MIC) and Total Information Coefficient (TIC) between causal representation $\mathbf{z}$ and $\mathbf{A} \odot \mathbf{u}$, where $\mathbf{A} = \mathbf{I}$ for all baseline models

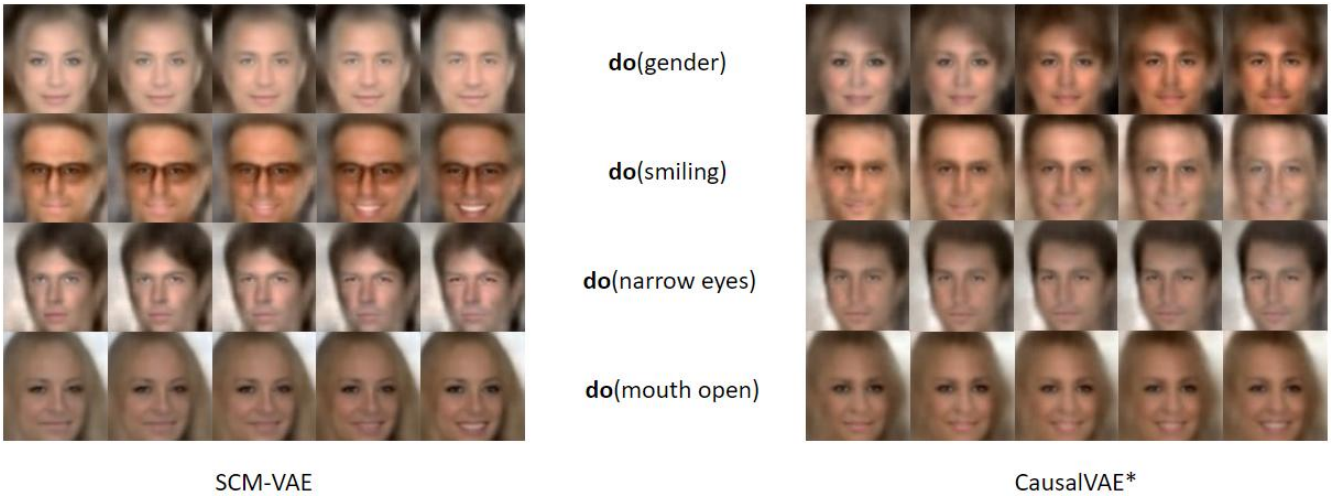

Figure 2. Intervention results for SCM-VAE (left) and CausalVAE (right) on CelebA-Smile dataset. We vary the value of the latent dimension in the range $(-1, 1)$ and observe changes to the image.

the known causal structure. The MIC and TIC values between the learned representation $\mathbf{z}$ and the causal labels are higher than CausalVAE in the same fixed causal graph setting on almost all datasets, as seen in Table I, which indicates that our choice of the prior is a more accurate representation of the ground-truth causal structure. Further, we observe that under do-intervention, our framework can achieve lower reconstruction loss compared to the ground-truth generated intervention on synthetic data, as shown in Tables II and III. Designing our prior such that the causal structure is encoded in the latent space enables our model to learn a posterior consistent with the causal graph. Thus, interventions on the causal concepts generate intervened representations decoded to be counterfactual instances.

We compare SCM-VAE to the CausalVAE baseline for evaluating interventions since it is the only VAE-based model capable of performing interventions. We perform the same atomic interventions on all test images. Under interventional queries, Figures 2, 3, and 4 show that our model yields more accurate interventions and can perform well given causal graphs compared to CausalVAE. For example, in Figure 3 (SCM-VAE), intervening on *pendulum angle* and *light position* accurately changes them to the desired values, respectively, and also changes the *shadow length* and *shadow position* accordingly. We can see that the CausalVAE intervention on *shadow position* also changes the *light position*, which means that extraneous causal dependencies are captured that are inconsistent with the modeling assumptions. This is likely because the causal mechanisms in CausalVAE cannot capture the non-linear relationship between causal variables since the causal structure is not explicitly encoded. However, in SCM-VAE, although interventions on leaf node variables are not always entirely accurate, the intervention leaves the parent variables unchanged, so the causes are independent of interventions on effect variables. Similarly, in Figure 4 (SCM-VAE), interventions on *ball size* change the water height appropriately, interventions on *hole* change the water flow only, and intervention on the leaf node *water flow* changes only the water flow, leaving other variables unchanged.

Since we do not have the ground-truth generating process for the CelebA images, we cannot evaluate the interventional reconstruction between the generated image and the ground-truth image. However, we can observe the quality of images generated under interventions. For instance, in the CelebA-

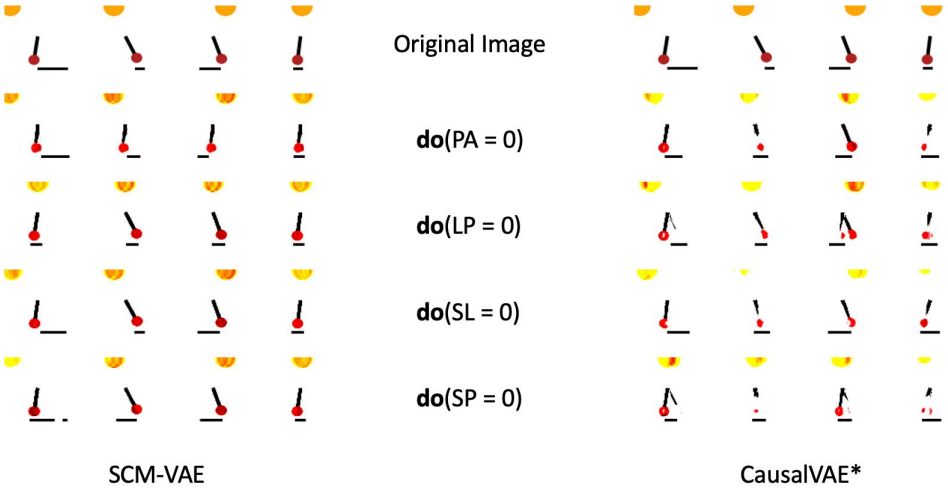

Figure 3. Intervention results for SCM-VAE (left) and CausalVAE (right) on Pendulum dataset where PA = pendulum angle, LP = light position, SL = shadow length, and SP = shadow position and the intervention is carried out via the $\mathbf{do}(\cdot)$ operator on the latent variable $\mathbf{z}$

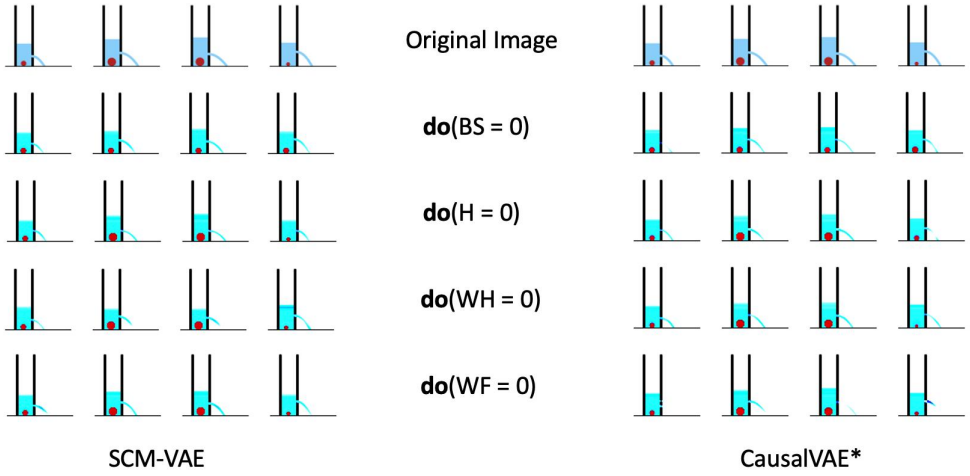

Figure 4. Intervention results for SCM-VAE (left) and CausalVAE (right) on Flow dataset where BS = ball size, H = hole, WH = water height, and WF = water flow and the intervention is carried out via the $\mathbf{do}(\cdot)$ operator on the latent variable $\mathbf{z}$

Smile example in Figure 2, intervening on the *smiling* causal concept in the representation space changes the person in the image to be smiling and changes the eye shape and mouth appearance accordingly. Further, intervention on leaf node variables such as *narrow eyes* only changes that concept and leaves other concepts unchanged. Note that by intervening on the *mouth open* concept, CausalVAE generates images where the *smiling* concept is also changed. SCM-VAE only changes the *mouth open* and the *narrow eyes* as consistent with the assumed causal graph. We can clearly see that our model can generate more realistic counterfactual instances under interventions compared to CausalVAE. Further, observe that interventions on a causal variable only affect its descendants and the generative factors are independent conditioned on their parents. The structural causal prior enforces this structure in the latent space.

	SCM-VAE	**CausalVAE***
do(*pendulum angle*)	**0.0221**±0.001	0.0637±0.002
do(*light position*)	**0.0185**±0.001	0.0587±0.003
do(*shadow length*)	**0.0281**±0.001	0.0389±0.002
do(*shadow position*)	**0.0273**±0.002	0.0412±0.004

TABLE II: Average Reconstruction Error on Post-Intervention Image (Pendulum dataset)

	SCM-VAE	**CausalVAE***
do(*ball size*)	**0.0289**±0.0003	0.0295±0.0003
do(*hole*)	0.0325±0.001	**0.0296**±0.0003
do(*water height*)	0.0316±0.0004	**0.0288**±0.0001
do(*water flow*)	**0.0639**±0.0004	0.0667±0.0001

TABLE III: Average Reconstruction Error on Post-Intervention Image (Flow dataset)

VII. Conclusion

In this work, we have studied learning causal representations using prior structural knowledge of high-level concepts. We proposed our framework, SCM-VAE, which utilizes a structural causal prior and a non-linear structural causal model formulation to learn the causal representation consistent with the assumed causal graph. We provided theoretical analysis and intuition on how the proposed prior can lead to learning accurate and identifiable causal representations. Further, our empirical evaluation showed that our model learns more consistent causal representations compared to baselines and is robust under interventions. In our future work, we will extend our framework to include efficient causal discovery methods and interventional data to learn causal representations.

VIII. Acknowledgement

This work is supported in part by National Science Foundation under awards 1946391 and 2147375, the National Institute of General Medical Sciences of National Institutes of Health under award P20GM139768, and the Arkansas Integrative Metabolic Research Center at University of Arkansas.

References

[1] J. Pearl, *Causality*, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[2] M. J. Vowels, N. C. Camgoz, and R. Bowden, "D'ya like dags? a survey on structure learning and causal discovery," *ACM Comput. Surv.*, 2022.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, p. 1798–1828, 2013.

[4] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[5] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "Causalvae: Disentangled representation learning via neural structural causal models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[6] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and S. Gavves, "CITRIS: Causal identifiability from temporal intervened sequences," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[7] C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf, "Invariant causal representation learning for out-of-distribution generalization," in *International Conference on Learning Representations*, 2022.

[8] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ica: A unifying framework," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

[9] S. Lachapelle, P. Rodriguez, Y. Sharma, K. E. Everett, R. L. PRIOL, A. Lacoste, and S. Lacoste-Julien, "Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA," in *First Conference on Causal Learning and Reasoning*, 2022.

[10] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen, "Weakly-supervised disentanglement without compromises," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[11] K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim, "The causal-neural connection: Expressiveness, learnability, and inference," in *Advances in Neural Information Processing Systems*, 2021.

[12] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[13] X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang, "Weakly supervised disentangled generative causal representation learning," *Journal of Machine Learning Research*, vol. 23, no. 241, pp. 1–55, 2022.

[14] I. Ng, S. Zhu, Z. Fang, H. Li, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," in *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, 2022.

[15] Z. Xu, J. Liu, D. Cheng, J. Li, L. Liu, and K. Wang, "Disentangled representation with causal constraints for counterfactual fairness," *arXiv:2208.09147 [cs, stat]*, 2022.

[16] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015.

[17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[18] P. Brakel and Y. Bengio, "Learning independent features with adversarial nets for non-linear ica," *arXiv:1710.05050 [cs, stat]*, 2017.

[19] R. Suter, D. o. Miladinovic, B. Schölkopf, and S. Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[20] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath, "Causalgan: Learning causal implicit generative models with adversarial training," *arXiv:1709.02023 [cs, stat]*, 2017.

[21] Y. Yu, J. Chen, T. Gao, and M. Yu, "Dag-gnn: Dag structure learning with graph neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[22] J. Brehmer, P. D. Haan, P. Lippe, and T. Cohen, "Weakly supervised causal representation learning," in *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

[23] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," in *Advances in Neural Information Processing Systems*, 2018.

[24] J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114 [cs, stat]*, 2013.

[26] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, *On Pearl's Hierarchy and the Foundations of Causal Inference*, 1st ed. Association for Computing Machinery, 2022, p. 507–556.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision*, 2015.

[28] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, 2014.

Appendix

A. Proof of Theorem 1

Proof. Applying Bayesian network factorization of $\mathbf{z}$ given $\mathbf{A}$ we obtain the following equation:

$$p(z_1, z_2, ..., z_n) = \prod_{i=1}^{n} p(z_i|\mathbf{Pa}(z_i))$$

Since for each i we have bijective mapping function $m_i \in \mathbf{m}$ such that $u_i = m_i(z_i)$, the causal relationships within variables $u_i, u_j \in \mathbf{u}$, $\forall i, j$ remains invariant under the bijective mapping between $\mathbf{z}$ and $\mathbf{u}$. Denote the causal graph for the representation vector $\mathbf{z}$ given $\mathbf{A}$ by $\mathcal{G}_{\mathbf{Z}}$. We can thus construct the causal graph $\mathcal{G}_{\mathbf{U}}$ for label vector $\mathbf{u}$ by inheriting all the cause-effect relationships in $\mathcal{G}_{\mathbf{Z}}$. Thus the Bayesian network factorization on $\mathbf{u}$ remains the same with $\mathcal{G}_{\mathbf{U}}$:

$$p(u_1, u_2, ..., u_n) = \prod_{i=1}^{n} p(u_i|\mathbf{Pa}(u_i))$$

where the corresponding functional mechanisms that determine each $p(u_i|\mathbf{Pa}(u_i))$ and $p(z_i|\mathbf{Pa}(z_i))$ are linked by the composition of bijective mapping functions $m_{i,\mathbf{pa}(i)} \in \mathbf{m}$. □

B. Proof of Theorem 2

Proof. **Step 1:** We first transform the equality of observed data distributions into equality of noiseless distributions. Suppose we have two sets of parameters $\boldsymbol{\theta} = (\mathbf{E}, \mathbf{D}, \mathbf{A}, \mathbf{T}, \boldsymbol{\lambda})$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{E}}, \tilde{\mathbf{D}}, \tilde{\mathbf{A}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$ such that $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u})$.

$$
\begin{aligned}
& p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u}) \\
\Rightarrow & \int \int_{\mathbf{z},\boldsymbol{\epsilon}} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z},\boldsymbol{\epsilon}) p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon},\mathbf{z}|\mathbf{u})\, d\mathbf{z} d\boldsymbol{\epsilon} \\
= & \int \int_{\mathbf{z},\boldsymbol{\epsilon}} p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}|\boldsymbol{\epsilon},\mathbf{z}) p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\epsilon},\mathbf{z}|\mathbf{u})\, d\boldsymbol{\epsilon} d\mathbf{z} \\
\Rightarrow & \int_{\mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})\, d\mathbf{z} = \int_{\mathbf{z}} p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{z}) p_{\tilde{\boldsymbol{\theta}}}(\mathbf{z}|\mathbf{u})\, d\mathbf{z} \\
\Rightarrow & \int_{\mathbf{x}'} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{D}^{-1}(\mathbf{x}')) p_{\boldsymbol{\theta}}(\mathbf{D}^{-1}(\mathbf{x}')|\mathbf{u})\, |\det(J_{\mathbf{D}^{-1}}(\mathbf{x}'))|\, d\mathbf{X}' \\
= & \int_{\mathbf{x}'} p_{\boldsymbol{\theta}}(\mathbf{x}|\tilde{\mathbf{D}}^{-1}(\mathbf{x}')) p_{\tilde{\boldsymbol{\theta}}}(\tilde{\mathbf{D}}^{-1}(\mathbf{x}')|\mathbf{u})\, |\det(J_{\tilde{\mathbf{D}}^{-1}}(\mathbf{x}'))|\, d\mathbf{X}'
\end{aligned}
\tag{20}
$$

Following assumption (1), the first term $p_{\boldsymbol{\theta}}(\mathbf{x}|\tilde{\mathbf{D}}^{-1}(\mathbf{x}')) = p_{\xi}(\mathbf{x} - \mathbf{x}')$ in the integral is vanished since the Gaussian distribution $p_{\xi}(\xi)$ could have infinitesimal variance. Further we have:

$$
\begin{aligned}
& p_{\boldsymbol{\theta}}(\mathbf{D}^{-1}(\mathbf{x}')|\mathbf{u})\, |\det(J_{\mathbf{D}^{-1}}(\mathbf{x}'))| \\
& \quad = p_{\tilde{\boldsymbol{\theta}}}(\tilde{\mathbf{D}}^{-1}(\mathbf{x}')|\mathbf{u})\, |\det(J_{\tilde{\mathbf{D}}^{-1}}(\mathbf{x}'))| \\
& \quad \Rightarrow \tilde{p}_{\boldsymbol{\theta}}(\mathbf{x}) = \tilde{p}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})
\end{aligned}
\tag{21}
$$

The result shown in Equation 21 implies that in order to make the marginal distribution invariant after adding noise, the noise-free deterministic distributions must be the same.

Step 2: We then construct $\boldsymbol{\lambda}_s(\mathbf{u}, \mathbf{A})$ following the definition of multivariate Gaussian distribution:

$$
\boldsymbol{\lambda}_s(\mathbf{u}, \mathbf{A}) = \begin{bmatrix} \lambda_s^1(\mathbf{u}, \mathbf{A}) & & \\ & \ddots & \\ & & \lambda_s^n(\mathbf{u}, \mathbf{A}) \end{bmatrix} \tag{22}
$$

Where $\lambda_s^i(\mathbf{u}, \mathbf{A}) = \lambda_s(u_i, \mathbf{Pa}_i(\mathbf{u}))$ for $i = 1, ..., n$ and s denotes the index of sufficient statistics of Gaussian distributions: $s = 1$ implies the mean and $s = 2$ implies the variance. By taking the logarithm on both side of the equation above we can derive the following equations:

$$
\begin{aligned}
& \log |\det(J_{\mathbf{D}^{-1}}(\mathbf{x}))| - \log \mathbf{Q}(\mathbf{D}^{-1}(\mathbf{x})) + \log \mathbf{Z}(\mathbf{u}) \\
& \qquad + \sum_{s=1}^{2} \mathbf{T}_s(\mathbf{D}^{-1}(\mathbf{x})) \boldsymbol{\lambda}_s(\mathbf{u}, \mathbf{A}) \\
= & \log |\det(J_{\tilde{\mathbf{E}}}(\mathbf{x}))| - \log \tilde{\mathbf{Q}}(\mathbf{D}^{-1}(\mathbf{x})) + \log \tilde{\mathbf{Z}}(\mathbf{u}) \\
& \qquad + \sum_{s=1}^{2} \tilde{\mathbf{T}}_s(\tilde{\mathbf{D}}^{-1}(\mathbf{x})) \tilde{\boldsymbol{\lambda}}_s(\mathbf{u}, \mathbf{A})
\end{aligned}
\tag{23}
$$

where $\mathbf{Q}$ denotes the base measure. Specifically in Gaussian distribution, it is $\boldsymbol{\sigma}(\mathbf{z})$. In the learning process, $\mathbf{A}$ is a full rank matrix and the items that are not related to u in the above equation are canceled out.

$$
\sum_{s=1}^{2} \mathbf{T}_s(\mathbf{D}^{-1}(\mathbf{x})) \boldsymbol{\lambda}_s(\mathbf{u}, \mathbf{A}) = \sum_{s=1}^{2} \tilde{\mathbf{T}}_s(\tilde{\mathbf{D}}^{-1}(\mathbf{x})) \tilde{\boldsymbol{\lambda}}_s(\mathbf{u}, \mathbf{A}) + \mathbf{b} \tag{24}
$$

where $\mathbf{b}$ is a vector related to $\mathbf{u}$.

The deterministic relationship between $\boldsymbol{\epsilon}$ and $\mathbf{z}$ could be expressed by $\mathbf{z} = \mathcal{T}(\boldsymbol{\epsilon}, \mathbf{A})$ via an invertible transformation function $\mathcal{T}$. We can thus derive the equivalent expression of Equation 24 with $\mathbf{D}^{-1} = \mathcal{T} \circ \mathbf{E}$, where $\circ$ represents function composition.

$$
\sum_{s=1}^{2} \mathbf{T}_s(\mathcal{T} \circ \mathbf{E}(\mathbf{x})) \boldsymbol{\lambda}_s(\mathbf{u}, \mathbf{A}) = \sum_{s=1}^{2} \tilde{\mathbf{T}}_s(\tilde{\mathcal{T}} \circ \tilde{\mathbf{E}}^{-1}(\mathbf{x})) \tilde{\boldsymbol{\lambda}}_s(\mathbf{u}, \mathbf{A}) + \mathbf{b} \tag{25}
$$

Step 3: We next construct an invertible matrix $\mathbf{L}$ corresponding to the label vector $\mathbf{u}$ and matrix $\mathbf{A}$:

$$
\mathbf{L} = \begin{bmatrix} \boldsymbol{\lambda}_1(\mathbf{u}, \mathbf{A}) & \\ & \boldsymbol{\lambda}_2(\mathbf{u}, \mathbf{A}) \end{bmatrix} \tag{26}
$$

According to the assumption that $\lambda_s^i(\mathbf{u}, \mathbf{A}) \neq 0, \forall i, s$, $\mathbf{L}$ is $2n \times 2n$ invertible and full rank diagonal matrix. Replacing the function of $\boldsymbol{\lambda}$ by $\mathbf{L}$ we could get:

$$
\mathbf{L}\mathbf{T}(\mathbf{D}^{-1}(\mathbf{x})) = \tilde{\mathbf{L}}\tilde{\mathbf{T}}(\tilde{\mathbf{D}}^{-1}(\mathbf{x})) + \mathbf{b} \tag{27}
$$

$$
\mathbf{T}(\mathbf{D}^{-1}(\mathbf{x})) = \mathbf{B}_2 \tilde{\mathbf{T}}(\tilde{\mathbf{D}}^{-1}(\mathbf{x})) + \mathbf{b}_2 \tag{28}
$$

where letting $r_{i,s} = \lambda_{i,s}(\mathbf{u}, \mathbf{A})$ and $\tilde{r}_{i,s} = \tilde{\lambda}_{i,s}(\mathbf{u}, \mathbf{A})$,

$$
\begin{aligned}
\mathbf{B}_2 &= \tilde{\mathbf{L}}/\mathbf{L} \\
&= \begin{bmatrix} r_{1,1}^{-1}\tilde{r}_{1,1} & & \\ & \ddots & \\ & & r_{n,2}^{-1}\tilde{r}_{n,2} \end{bmatrix}
\end{aligned}
\tag{29}
$$

We then replace $\mathbf{D}^{-1}$ with $\mathcal{T} \circ \mathbf{E}$ and derive the following equation:

$$
\begin{aligned}
& \mathbf{L}\mathbf{T}(\mathcal{T} \circ \mathbf{E}(\mathbf{x})) = \tilde{\mathbf{L}}\tilde{\mathbf{T}}(\tilde{\mathcal{T}} \circ \tilde{\mathbf{E}}(\mathbf{x})) \\
& \Rightarrow \mathbf{T}(\mathbf{E}(\mathbf{x})) = \mathbf{B}_1 \tilde{\mathbf{T}}(\tilde{\mathbf{E}}(\mathbf{x})) + \mathbf{b}_1
\end{aligned}
\tag{30}
$$

Next we adopt the proof idea in [5] to show both $\mathbf{B}_1$ and $\mathbf{B}_2$ are invertible matrices. Following Lemma 3 in [8], we are able to select a pair $(\epsilon_i, \epsilon_i^2)$ to make $(\mathbf{T}_i'(z_i), \mathbf{T}_i'(z_i^2))$ linearly independent. Then we concat the two points into a vector and denote the Jacobian matrix $\mathbf{Q} = [J_{\mathbf{T}}(\epsilon), J_{\mathbf{T}}(\epsilon^2)]$, and define $\tilde{\mathbf{Q}}$ on $\tilde{\mathbf{T}}(\tilde{\mathbf{E}} \circ \mathcal{T} \circ \mathbf{D}(\boldsymbol{\epsilon}))$ in the same manner. By differentiating Equation 30 we have

$$
\mathbf{Q} = \mathbf{B}_1 \tilde{\mathbf{Q}} \tag{31}
$$

According to assumption (2) the Jacobian matrix of $\mathbf{D}^{-1}$ is full rank. Thus both $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$ are invertible matrices. From Equation 31 we can derive $\mathbf{B}_1$ is also an invertible matrix. By applying similar procedure we can prove the invertibility of $\mathbf{B}_2$ as well. Finally, the invertibility of $\mathbf{B}_1$ and $\mathbf{B}_2$ leads to $\sim$ identifiability of our model parameters. □