CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models

Mengyue Yang[1], **Furui Liu**[1], **Zhitang Chen**[1], **Xinwei Shen**[2], **Jianye Hao**[1], **Jun Wang**[3]
[1] Noah's Ark Lab, Huawei, Shenzhen, China
[2] The Hong Kong University of Science and Technology, Hong Kong, China
[3] University College London, London, United Kingdom
yangmengyue2,liufurui2,chenzhitang2,haojianye@huawei.com
xshenal@connect.ust.hk
jun.wang@cs.ucl.ac.uk

Abstract

Learning disentanglement aims at finding a low dimensional representation which consists of multiple explanatory and generative factors of the observational data. The framework of variational autoencoder (VAE) is commonly used to disentangle independent factors from observations. However, in real scenarios, factors with semantics are not necessarily independent. Instead, there might be an underlying causal structure which renders these factors dependent. We thus propose a new VAE based framework named CausalVAE, which includes a Causal Layer to transform independent exogenous factors into causal endogenous ones that correspond to causally related concepts in data. We further analyze the model identifiabitily, showing that the proposed model learned from observations recovers the true one up to a certain degree. Experiments are conducted on various datasets, including synthetic and real word benchmark CelebA. Results show that the causal representations learned by CausalVAE are semantically interpretable, and their causal relationship as a Directed Acyclic Graph (DAG) is identified with good accuracy. Furthermore, we demonstrate that the proposed CausalVAE model is able to generate counterfactual data through "do-operation" to the causal factors.

1 Introduction

Disentangled representation learning is of great importance in various applications such as computer vision, speech and natural language processing, and recommender systems Hsu et al. [2017], Ma et al. [2019], Hsieh et al. [2018]. The reason is that it might help enhance the performance of models, i.e. improving the generalizability, robustness against adversarial attacks as well as the explanability, by learning data's latent disentangled representation. One of the most common frameworks for disentangled representation learning is Variational Autoencoders (VAE), a deep generative model trained to disentangle the underlying explanatory factors. Disentanglement via VAE can be achieved by a regularization term of the Kullback-Leibler (KL) divergence between the posterior of the latent factors and a standard Multivariate Gaussian prior, which enforces the learned latent factors to be as independent as possible. It is expected to recover the latent variables if the observation in real world is generated by countable independent factors. To further enhance the independence, various extensions of VAE consider minimizing the mutual information among latent factors. For example, Higgins et al. [2017] and Burgess et al. [2018] increased the weight of the KL divergence term to enforce independence. Kim and Mnih [2018], Chen et al. [2018] further encourage the independence by reducing total correlation among factors.

Most existing works of disentangled representation learning make a common assumption that the real world observations are generated by countable independent factors. Nevertheless we argue that in

Preprint. Under review.

Figure 1: A swinging pendulum: an illustrative example

many real world applications, latent factors with semantics of interest are causally related and thus we need a new framework that supports causal disentanglement.

Consider a toy example of a swinging pendulum in Fig. 1. The position of the illumination source and the angle of the pendulum are causes of the position and the length of the shadow. Through causal disentangled representation learning, we aim at learning representations that correspond to the above four concepts. Obviously, these concepts are not independent and existing methods may fail to extract those factors. Furthermore, causal disentanglement allow us to manipulate the causal system to generate counterfactual data. For example, we can manipulate the latent code of shadow to create new pictures without shadow even there are pendulum and light. This corresponds to the "do-operation" Pearl [2009] in causality, where the system operates under the condition that certain variables are controlled by external forces. A deep generative model that supports "do-operation" is of tremendous value as it allows us to ask “what-if" questions when making decisions.

In this paper, we propose a VAE-based causal disentangled representation learning framework by introducing a novel Structural Causal Model layer, which allows us to recover the latent factors with semantics and structured via a causal DAG. The input signal passes through an encoder to obtain independent exogenous factors and then a Causal Layer to generate causal representation which is taken by the decoder to reconstruct the original input. We call the whole process Causal Disentangled Representation Learning. Unlike unsupervised disentangled representation learning of which the feasibility is questionable Locatello et al. [2018], additional information is required as weak supervision signals to achieve causal representation learning. By “weak supervision", we emphasize that in our work, the causal structure of the latent factors is automatically learned, instead of being given as a prior in Kocaoglu et al. [2017]. To train our model, we propose a new loss function which includes the VAE evidence lower bound loss and an acyclicity constraint imposed on the learned causal graph to guarantee its “DAGness". In addition, we analyze the identifiablilty of the proposed model, showing that the learned parameters of the disentangled model recover the true one up to certain degree. The contribution of our paper is three-fold. (1) We propose a new framework named CausalVAE that supports causal disentanglement and “do-operation"; (2) Theoretical justification on model identifiability is provided; (3) We conduct comprehensive experiments with synthetic and real world face images to demonstrate that the learned factors are with causal semantics and can be intervened to generate counterfactual images that do not appear in training data.

2 Related Works

In this section, we review state-of-the-art disentangled representation learning methods, including some recent advances on combining causality and disentangled representation learning. We also present preliminaries of causal structure learning from pure observations which is a key ingredient of our proposed CausalVAE framework.

Disentangled Representation Learning: Conventional disentangled representation learning methods learn mutually independent latent factors by an encoder-decoder framework. In this process, a standard normal distribution is used as a prior of the latent code. A variational posterior $q(\mathbf{z}|\mathbf{x})$ is then used to approximate the unknown true posterior $p(\mathbf{z}|\mathbf{x})$. This framework was further extended by adding new independence regularization terms to the original loss function, leading to various algorithms. β-VAE Higgins et al. [2017] proposes an adaptation framework which adjusts the weight of KL term to balance between independence of disentangled factors and the reconstruction performance. While factor VAE Chen et al. [2018] proposes a new framework which focuses solely on the independence of factors. Ladder VAE Lee et al. [2016] on the other hand, leverages the structure of ladder neural network to train a structured VAE for hierarchical disentanglement. Nevertheless the aforementioned unsupervised disentangled representation learning algorithms do not perform well in some situations where there is complex causal relationship among factors. Furthermore, they are challenged for lacking inductive bias and thus the model identifiability cannot be guaranteed

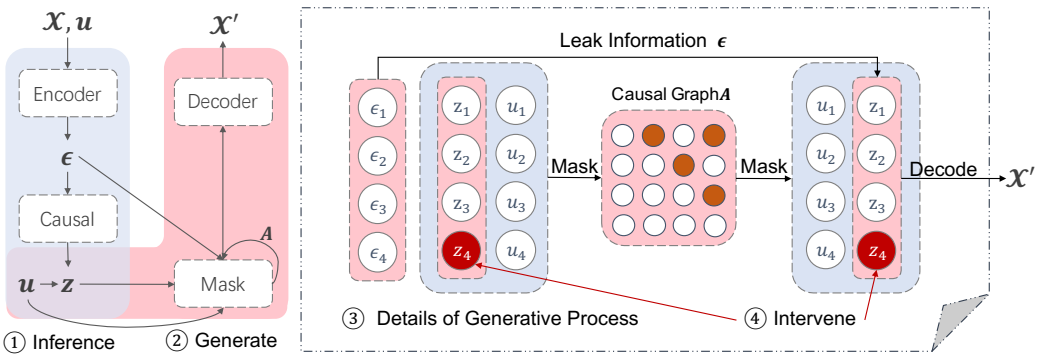


Figure 2: Model structure of CausalVAE. The encoder takes observation $\mathbf{x}$ as inputs to generate independent exogenous variable $\boldsymbol{\epsilon}$, whose prior distribution is assumed to be standard Multivariate Gaussian. Then it is transformed by the Causal Layer into causal representations $\mathbf{z}$ (Eq. 1) with a conditional prior distribution $p(\mathbf{z}|\mathbf{u})$. A Mask Layer is then applied to $\mathbf{z}$ to resemble the SCM in Eq. 2. After that, $\mathbf{z}$ is taken as the input of the decoder to reconstruct the observation $\mathbf{x}$.

Locatello et al. [2018]. The identifiability problem of VAE is defined as follows: if the parameters $\tilde{\boldsymbol{\theta}}$ learned from data lead to a marginal distribution equal to the true one parameterized by $\boldsymbol{\theta}$, i.e., $p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x})$, then the joint distributions also match, i.e. $p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$. Therefore, the rotation invariance of prior $p(\mathbf{z})$ (standard Multivariate Gaussian distribution) will lead the unindentifiable of $p(\mathbf{z})$. Khemakhem et al. [2019] prove that there is infinite number of distinct models entailing the same joint distributions, which means that the underlying generative model is not identifiable through unsupervised learning. On the contrary, by leveraging a few labels, one is able to recover the true model Mathieu et al. [2018], Locatello et al. [2018]. Kulkarni et al. [2015] and Locatello et al. [2019] use additional labels to reduce the model ambiguity. Khemakhem et al. [2019] gives an identifiability of VAE with additional inputs, by leveraging the theory of nonlinear Independent Component Analysis (nonlinear ICA) Brakel and Bengio [2017].

Causal Discovery & Causal Disentangled Representation Learning: We refer to causal representation as ones structured by a causal graph. Discovering the causal graph from pure observations has attracted large amounts of attention in the past decades Hoyer et al. [2009], Zhang and Hyvarinen [2012], Shimizu et al. [2006]. Pearl [2009] introduced a Probabilistic Graphical Models (PGMs) based language to describe causality among variables. Shimizu et al. [2006] proposed an effective method called LiNGAM to learn the causal graph and they prove the model identifiability under the linearity and non-Gaussianity assumption. Zheng et al. [2018] proposed NOTEARs with a fully differentiable DAG constraint for causal structure learning, which drastically reduces a very complicated combinatorial optimization problem to a continuous optimization problem. Zhu et al. [2020] proposed a flexible and efficient Reinforcement Learning (RL) based method to search over a DAG space for a best graph with a highest score. Recently, the community has raised interest of combining causality and disentangled representation. Suter et al. [2018] used causality to explain disentangled latent representations. Kocaoglu et al. [2017] proposed a method called CausalGAN which supports "do-operation" on images but it requires the causal graph given as a prior. Instead of assuming independent latent factors, Besserve et al. [2018] allows dependent latent factors. However, in their work, the dependence is induced by some latent confounders, instead of a causal graph among latent factors investigated in this paper. Schölkopf [2019] stressed the importance and necessity of causal disentangled representation learning but it still remains conceptual. To the best of our knowledge, our work is the first one that successfully implements the idea of causal disentanglement.

3 Causal Disentanglement in Variational Autoencoder

We start with the definition of causal representation, and then propose a new framework to achieve causal disentanglement by leveraging additional inputs, e.g. labels of concepts. Firstly, we give an overview of our proposed CausalVAE model structure in Fig. 2. A Causal Layer, which essentially describes a Structural Causal Model (SCM) Shimizu et al. [2006], is introduced to a conventional VAE network. The Causal Layer transforms the independent exogenous factors to causal endogenous factors corresponding to causally related concepts of interest. A mask mechanism Ng et al. [2019a] is then used to propagate the effect of parental variables to their children, mimicking the assignment

operation of SCMs. Such a Causal Layer is the key to supporting intervention or "do-operation" to the system.

3.1 Transforming Independent Exogenous Factors into Causal Representations

Our model is within the framework of VAE-based disentanglement. In addition to the encoder and the decoder structures, we introduce a Structural Causal Model (SCM) layer to learn causal representations. To formalize causal representation, we consider n concepts of interest in data. The concepts in observations are causally structured by a Directed Acyclic Graph (DAG) with an adjacency matrix $\mathbf{A}$. Though a general nonlinear SCM is preferred, for simplicity, in this work, the Causal Layer exactly implements a Linear SCM as described in Eq. 1 (shown in Fig. 2 ①),

$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (I - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\mathbf{A}$ is the parameters to be learnt in this layer. $\boldsymbol{\epsilon}$ are independent Gaussian exogenous factors and $\mathbf{z} \in \mathbb{R}^n$ is structured causal representation of n concepts that is generated by a DAG and thus $\mathbf{A}$ can be permuted into a strictly upper triangular matrix.

Unsupervised learning of the model might be infeasible due to the identifiability issue as discussed in Locatello et al. [2018]. To address this problem, similar to iVAE Khemakhem et al. [2019], we adopt additional information $\mathbf{u}$ associated with the true causal concepts as supervising signals. In our work, we use the labels of the concepts. The additional information $\mathbf{u}$ is utilized in two ways. Firstly, we propose a conditional prior $p(\mathbf{z}|\mathbf{u})$ to regularize the learned posterior of $\mathbf{z}$. This guarantees that the learned model belongs to an identifiable family. Secondly, we also leverage $\mathbf{u}$ to learn the causal structure $\mathbf{A}$. Besides learning the causal representations, we further enable the model to support intervention to the causal system to generate counterfactual data which does not exist in the training data.

3.2 Structural Causal Model Layer

Once the causal representations $\mathbf{z}$ is obtained, it passes through a Mask Layer Ng et al. [2019a] to reconstruct itself. Note that this step resembles a SCM which depicts how children are generated by their corresponding parental variables. We will show why such a layer is necessary to achieve intervention. Let z_i be the ith variable in the vector $\mathbf{z}$. The adjacency matrix associated with the causal graph is $\mathbf{A} = [\mathbf{A}_1|\dots|\mathbf{A}_n]$ where $\mathbf{A}_i \in \mathbb{R}^n$ is the weight vector such that A_{ji} encodes the causal strength from z_j to z_i. We have a set of mild nonlinear and invertible functions $[g_1, g_2, \dots, g_n]$ that map parental variables to the child variable. Then we write

$$z_i = g_i(\mathbf{A}_i \circ \mathbf{z}; \boldsymbol{\eta}_i) + \epsilon_i, \tag{2}$$

where $\circ$ is the element-wise multiplication and $\boldsymbol{\eta}_i$ is the parameter of $g_i(\cdot)$ (as shown in Fig. 2 ③). Note that according to Eq. 1, we can simply write $z_i = \mathbf{A}_i^T\mathbf{z} + \epsilon_i$. However, we find that adding a mild nonlinear function g_i results in more stable performances. To show how this masking works, consider a variable z_i and $\mathbf{A}_i \circ \mathbf{z}$ equals a vector that only contains its parental information as it masks out all z_i's non-parent variables. By minimizing the reconstruction error, the adjacency matrix $\mathbf{A}$ and the parameter $\boldsymbol{\eta}_i$ of the mild nonlinear function g_i are trained.

This layer makes intervention or "do-operation" possible. Intervention Pearl [2009] in causality refers to modifying a certain part of a system by external forces and one is interested in the outcome of such manipulation. To intervene z_i, we set z_i on the RHS of Eq. 2 (corresponding to the $i-$th node of $\mathbf{z}$ in the first layer in Fig. 2) to a fixed value, and then its effect is delivered to all its children as well as itself on the LHS of Eq. 2 (corresponding to some nodes of $\mathbf{z}$ in the second layer). Note that intervening the cause will change the effect, whereas intervening the effect, on the other hand, does not change the cause because information can only flow into the next layer from the previous one in our model, which is aligned with the definition of causal effects.

3.3 A Probabilistic Generarive Model for CausalVAE

We give a probabilistic formulation of the proposed generative model (shown in Fig. 2 ②). Denote by $\mathbf{x} \in \mathbb{R}^d$ the observed variables and $\mathbf{u} \in \mathbb{R}^n$ the additional information. u_i is the label of the i-th concept of interest in data. Let $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be the latent exogenous independent variables and $\mathbf{z} \in \mathbb{R}^n$ be the latent endogenous variables with semantics where $\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon}$. For simplicity, we denote $\mathbf{C} = (\mathbf{I} - \mathbf{A}^T)^{-1}$.

We treat both $\mathbf{z}$ and $\boldsymbol{\epsilon}$ as latent variables. Consider the following conditional generative model parameterized by $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{h}, \mathbf{C}, \mathbf{T}, \boldsymbol{\lambda})$:

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

Let $\mathbf{f}(\mathbf{z})$ denote the decoder which is assumed to be an invertible function and $\mathbf{h}(\mathbf{x}, \mathbf{u})$ denotes the encoder. We define the generative and inference models as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \equiv p_{\boldsymbol{\xi}}(\mathbf{x} - \mathbf{f}(\mathbf{z})), \quad q_{\boldsymbol{\phi}}(\mathbf{z}, \boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u}) \equiv q(\mathbf{z}|\boldsymbol{\epsilon}) q_{\boldsymbol{\zeta}}(\boldsymbol{\epsilon} - \mathbf{h}(\mathbf{x}, \mathbf{u})), \tag{4}$$

which is obtained by assuming the following decoding and encoding processes:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\xi}, \quad \boldsymbol{\epsilon} = \mathbf{h}(\mathbf{x}, \mathbf{u}) + \boldsymbol{\zeta}, \tag{5}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are the vectors of independent noise with probability densities $p_{\boldsymbol{\xi}}$ and $q_{\boldsymbol{\zeta}}$. When $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are infinitesimal, the encoder and decoder can be regarded as deterministic ones. We define the joint prior $p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})$ for latent variables $\mathbf{z}$ and $\boldsymbol{\epsilon}$ as

$$p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}), \tag{6}$$

where $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the prior of latent endogenous variables $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is a factorized Gaussian distribution conditioning on the additional observation $\mathbf{u}$, i.e.

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i)), \tag{7}$$

where λ_1 and λ_2 are an arbitrary functions. In this paper, we let $\lambda_1(\mathbf{u}) = \mathbf{u}$ and $\lambda_2(\boldsymbol{u}) \equiv 1$.

4 Learning Strategy

In this section, we discuss how to train the CausalVAE model in order to learn the causal representation as well as the causal graph simultaneously.

4.1 Evidence Lower Bound of CausalVAE

We apply variational Bayes to learn a tractable distribution $q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})$ to approximate the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})$. Given data set $\mathcal{X}$ with the empirical data distribution $q_{\mathcal{X}}(\mathbf{x}, \mathbf{u})$, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are learned by optimizing the following evidence lower bound (ELBO):

$$\mathbb{E}_{q_{\mathcal{X}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})] \geq \text{ELBO} = \mathbb{E}_{q_{\mathcal{X}}}[\mathbb{E}_{\boldsymbol{\epsilon}, \mathbf{z} \sim q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u})] - \mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}))], \tag{8}$$

where $\mathcal{D}(\cdot\|\cdot)$ denotes KL divergence. Eq. 8 is intractable in general. However, thanks to the one-to-one correspondence between $\boldsymbol{\epsilon}$ and $\mathbf{z}$, we simplify the variational posterior as follows:

$$q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) = q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u}) \delta(\boldsymbol{z} = \mathbf{C}\boldsymbol{\epsilon}) = q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) \delta(\boldsymbol{\epsilon} = \mathbf{C}^{-1}\boldsymbol{z}), \tag{9}$$

where $\delta(\cdot)$ is the Dirac delta function. According to the model assumptions introduced in Section 3.3, i.e., generation process (Eq. 4) and prior (Eq. 6), we attain a neat form of ELBO loss as follows:

Proposition 1. ELBO defined in Eq. 8 can be written as:

$$\text{ELBO} = \mathbb{E}_{q_{\mathcal{X}}}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \mathcal{D}(q_\phi(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})) - \mathcal{D}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}))]. \tag{10}$$

Details of the proof are given in the Appendix. With this form, we can easily implement a loss function to train the CausalVAE model.

4.2 Learning the Causal Structure of Latent Codes

In addition to the encoder and decoder, our CausalVAE model involves a Causal Layer with a DAG structure to be learned. Note that both $\mathbf{z}$ and $\mathbf{A}$ are unknown, to ease the training task, we leverage the additional labels $\mathbf{u}$ to construct the following constraint:

$$l_u = \mathbb{E}_{q_{\mathcal{X}}} \|\mathbf{u} - \sigma(\mathbf{A}^T \mathbf{u})\|_2^2 \leq \kappa_1, \tag{11}$$

where σ is a logistic function as our lables are binary and κ_1 is the small positive constant value. This follows the idea that $\mathbf{A}$ should also describe the causal relations among labels well. Similarly we apply the same constraint to the learned latent code $\mathbf{z}$ as follows:

$$l_m = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi}}} \sum_{i=1}^{n} \|z_i - g_i(\mathbf{A}_i \circ \mathbf{z}; \boldsymbol{\eta}_i)\|^2 \leq \kappa_2, \tag{12}$$

where κ_2 is the small positive constant value. Lastly, the causal adjacency matrix $\mathbf{A}$ is constrained to be a DAG. Instead of using traditional DAG constraint that is combinatorial, we adopt a continuous differentiable constraint function Zheng et al. [2018], Zhu and Chen [2019], Ng et al. [2019b], Yu et al. [2019] . The function attains 0 if and only if the adjacency matrix $\mathbf{A}$ corresponds to a DAG Yu et al. [2019], i.e.

$$H(\mathbf{A}) \equiv tr((\mathbf{I} + \mathbf{A} \circ \mathbf{A})^n) - n = 0. \tag{13}$$

The training procedure of our CausalVAE model reduces to the following constrained optimization:

$$\text{maximize} \quad \text{ELBO}, \quad \text{s.t.} \quad (11)(12)(13).$$

By lagrangian multiplier method, we have the new loss function

$$\mathcal{L} = -\text{ELBO} + \alpha H(\mathbf{A}) + \beta l_u + \gamma l_m, \tag{14}$$

where α, β, γ denote regularization hyperparameters.

5 Identifiability Analysis

In this section, we present the identifiability of our proposed model. We adopt the $\sim$-*identifiability* [Khemakhem et al., 2019] as follows:

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

$$\begin{aligned} (\mathbf{f}, \mathbf{h}, \mathbf{C}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\tilde{\mathbf{f}}, \tilde{\mathbf{h}}, \tilde{\mathbf{C}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}) \Leftrightarrow \exists \mathbf{B_1}, \mathbf{B_2}, \mathbf{b_1}, \mathbf{b_2} | \\ \mathbf{T}(\mathbf{h}(\mathbf{x}, \mathbf{u})) = \mathbf{B_1}\tilde{\mathbf{T}}(\tilde{\mathbf{h}}(\mathbf{x}, \mathbf{u})) + \mathbf{b_1}, \mathbf{T}(\mathbf{f^{-1}}(\mathbf{x})) = \mathbf{B_2}\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{\mathbf{-1}}(\mathbf{x})) + \mathbf{b_2}, \forall \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{15}$$

where $\mathbf{C} = (\mathbf{I} - \mathbf{A}^T)^{-1}$. If $\mathbf{B_1}$ is an invertible matrix and $\mathbf{B_2}$ is an invertible diagonal matrix with diagonal elements associated to u_i. We say that the model parameter is $\sim$-*identifiable*. Following Khemakhem et al. [2019], we obtain the identifiability of our causal generative model as follows.

Theorem 1. Assume that the data we observed are generated according Eq. 3-4 and the following assumptions hold,

1. The set $\{x \in \mathcal{X} | \phi_\xi(\mathbf{x}) = 0\}$ has measure zero, where ϕ_ξ is the characteristic function of the density p_ξ defined in Eq. 5.
2. The decoder function $\mathbf{f}$ is differentiable and the Jacobian matrix of $\mathbf{f}$ is of full rank [1].
3. The sufficient statistics $T_{i,s}(z_i) \neq 0$ almost everywhere for all $1 \leq i \leq n$ and $1 \leq s \leq 2$, where $T_{i,s}(z_i)$ is the sth statistic of variable z_i.
4. The additional observations $u_i \neq 0$.

Then the parameters $(\mathbf{f}, \mathbf{h}, \mathbf{C}, \mathbf{T}, \boldsymbol{\lambda})$ are $\sim$-*identifiable*.

The identifiability of the model under supervision of additional information is obtained thanks to the conditional prior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$. The conditional prior guarantees that sufficient statistics of $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ are related to the value of $\mathbf{u}$. A complete proof of **Theorem 1** is available in Appendix.

6 Experiments

In this section, we conduct experiments using both synthetic dataset and real human face image dataset and we compare our CausalVAE model against existing state of the art methods on disentangled representation learning. We focus on examing whether a certain algorithm is able to learn interpretable representations and whether outcomes of intervention on learned latent code is consistent to our understanding of the causal system.

6.1 Dataset, Baselines & Metrics

Dataset: We conduct experiments on a synthetic dataset and a benchmark face dataset CelebA. The synthetic one is named Pendulum which includes images of causally related objects. Each image contains 3 entities (PENDULUM, LIGHT, SHADOW), and 4 concepts ((PENDULUM ANGLE, LIGHT

[1] (rank equals to its smaller dimension)

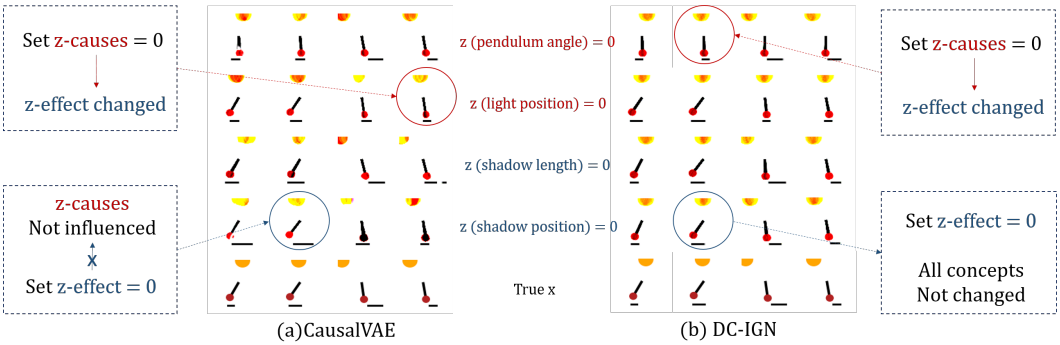


Figure 3: The results of Intervention experiments on the pendulum dataset. Each row shows the result of controlling the PENDULUM ANGLE, LIGHT ANGLE, SHADOW LENGTH, and SHADOW LOCATION respectively. The bottom row is the original input image.

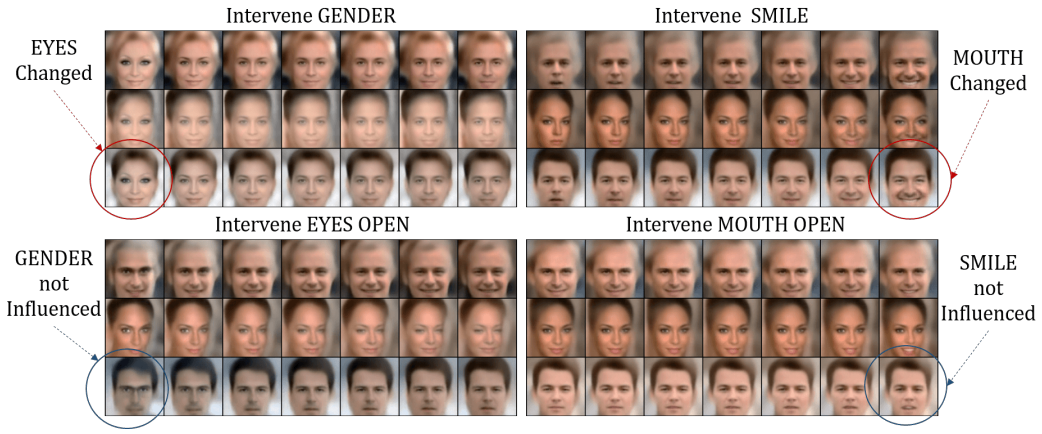


Figure 4: Results of CausalVAE model on CelebA. The controlled factors are GENDER, SMILE, EYES OPEN and MOUTH OPEN respectively.

ANGLE) $\rightarrow$ (SHADOW LOCATION, SHADOW LENGTH)). We also use a real world dataset CelebA[2], a widely used dataset in the computer vision community. In this dataset, there are in total 200k human face images with labels on different concepts, and we focus on 4 causally related concepts (GENDER, SMILE, EYES OPEN,MOUTH OPEN), where GENDER and SMILE cause EYES OPEN, and SMILE causes MOUTH OPEN. More experimental results on other concepts are provided in the Appendix.

Baselines: We compare our method with some state of the arts. They are categorized into supervised and unsupervised methods. CausalVAE-unsup, LadderVAE Lee et al. [2016] and β-VAE Higgins et al. [2017] are unsupervised methods. CausalVAE-unsup is a reduced version of our model whose structure is the same as CausalVAE except that the Mask Layer and the supervision conditional prior $p(\mathbf{z}|\mathbf{u})$ are removed. Supervised methods include disentangled representation learning method DC-IGN Kulkarni et al. [2015] and causal generative model CausalGAN Kocaoglu et al. [2017]. As CausalGAN does not focus on representation learning, we only compare our CausalVAE with CausalGAN on intervention experiment (results given in Appendix). For these methods, the prior conditioning on the labels are given, and the dimensionality of the latent representation is the same as CausalVAE.

Metrics: We use Maximal Information Coefficient (MIC) and Total Information Coefficient (TIC) Kinney and Atwal [2014] as our evaluation metrics. Both of them indicate the degree of information relevance between the learned representation and the ground truth labels of concepts.

[2]http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

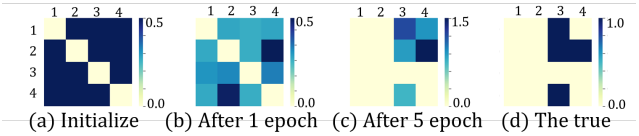


Figure 5: The learning process of causal matrix $\mathbf{A}$. The concepts include: GENDER, SMILE, EYES OPEN, MOUTH OPEN (top-to-bottom and left-to-right order); (c) converged $\mathbf{A}$, (d) ground truth .

Table 1: The MIC and TIC between learned representation $\mathbf{z}$ and the label $\mathbf{u}$. The results show that among all compared methods, the learned factors of our proposed CausalVAE achieve best alignment to the concepts of interest. (Note: the metrics include mean $\pm$ standard errors in table.)

	CausalVAE		DC-IGN		β-VAE		CausalVAE-unsup		LadderVAE	
Metrics(%)	MIC	TIC	MIC	TIC	MIC	TIC	MIC	TIC	MIC	TIC
Pendulum	**96.3**$_{\pm 3.6}$	**89.0**$_{\pm 2.9}$	61.8$_{\pm 8.7}$	48.1$_{\pm 7.3}$	22.6$_{\pm 4.6}$	12.5$_{\pm 2.2}$	21.2$_{\pm 1.4}$	12.0$_{\pm 1.0}$	22.4$_{\pm 3.1}$	12.8$_{\pm 1.2}$
CelebA	**83.7**$_{\pm 6.2}$	**71.6**$_{\pm 7.2}$	78.8$_{\pm 10.9}$	66.1$_{\pm 12.1}$	22.5$_{\pm 1.2}$	9.92$_{\pm 1.2}$	27.2$_{\pm 5.3}$	14.6$_{\pm 4.2}$	23.5$_{\pm 3.0}$	10.3$_{\pm 1.6}$

6.2 Intervention experiments

Intervention experiments aim at testing if a certain dimension of the latent representation has interpretable semantics. The value of a latent code is manipulated by "do-operation" as introduced in previous sections, and we observe how the generated image appears. Intervention is conducted by the following steps: 1) a generative model is trained; 2) an arbitrary image from the training set is fed to the encoder to generate a latent code $\mathbf{z}$. 3) we manipulate the value of z_i corresponding to a concept of interest. For CausalVAE, as Fig. 2 ④ shows, we need to manipulate both the input and output nodes of the SCM layer. Note that the effect of manipulation to a parental node will be propagated to its children; 4) The intervened latent code $\tilde{\mathbf{z}}$ passes through the decoder to generate a new image. In the experiments, all images in the dataset are used to train our proposed model CausalVAE and other baselines. Parameters $(\alpha, \beta, \gamma) = (1, 1, 1)$ for all experiments unless specified.

We first conduct intervention experiments on the Pendulum dataset, with 4 latent concepts and results are given in Fig. 3. We intervene a certain concept by setting the corresponding latent code value to 0. We expect that the pattern of the manipulated concept will be fixed across all images under the same intervention. For example, when we intervene the pendulum ANGLE as shown in the first line of Fig. 3 (a), the ANGLE of pendulum of different images are almost the same. Meanwhile, we also observe that the SHADOW LOCATION and SHADOW LENGTH change in a correct way that aligns with the physics law. Note that this is also related to the concept of modularity, meaning that intervening a certain part of the generative system usually does not affect the other parts of the system. Similar phenomenon is observed in other intervention experiments, demonstrating that our model correctly implement the underlying causal system. The results of DC-IGN, a supervised method without considering the causal structure, are given in Fig. 3 (b). There exists a problem that manipulating the latent codes of effects sometimes has no influence to the whole image. This is probably because they do not explicitly consider causal disentanglement.

Fig. 4 demonstrates the good result of CausalVAE on real world banchmark dataset CelebA, with subfigures showing the experiments on intervening concepts GENDER, SMILE, EYES OPEN and MOUTH OPEN respectively. We observe that when we intervene the cause concept SMILE, the status of MOUTH OPEN also changes. In contrast, intervening effect concept MOUTH OPEN does not cause the cause concept SMILE to change. Table 1 records the mutual information (MIC/TIC) between the learned representation and the ground truth concept labels of all compared methods. Our model achieves best alignment with the concept labels, justifying the effectiveness of our proposed method. On the contrary, factors learned by those compared methods have low correlation with the ground truth labels, indicating that those factors are at least not corresponding to the causal concepts of interest. In addition, we show in Fig. 5 the learned adjacency matrix $\mathbf{A}$. As the training epoch increases, we see that the graph learned by our model quickly converges to the true one, which shows that our method is able to correctly learn the causal relationship among the factors.

7 Conclusion

In this paper, we investigate an important task of learning disentangled representations of causally related concepts in data, and propose a new framework called CausalVAE which includes a SCM

layer to model the causal generation mechanism of data. We prove that the proposed model is fully identifiability given additional supervision signal. Experimental results with synthetic and real data show that CausalVAE successfully learns representations of causally related concepts and allows intervention to generate counterfactual outputs as expected according to our understanding of the causal system. To the best of our knowledge, our work is the first one that successfully implement causal disentanglement and is expected to bring new insights into the domain of disentangled representation learning.

Broader Impact

The proposed CausalVAE is a structured disentanglement representation learning method, which learns the low dimensional representation that aligns to the causally related concepts. The method creates a new branch of disentanglement learning, and the factors are structured with interpretability.

Possible Application Scenarios: Because the factors are able to be manipulated, our method is applicable to scenarios such as image and video understanding and modification, autonomous driving tasks. It could clear the factors like shadows on road to reduce the predict error and enhance the safety of autonomous driving.

The disentangled representation might be helpful in various downstream tasks, like clustering and classification, where the geometry of the original data is reshaped in low dimensional representation space.

The model learned from observations is applicable to build simulators for various scenarios where the system is decomposable into causal factors. The simulators can model counterfactual situations, allowing intervention operation and in-depth analysis of the behaviour of the system.

Possible Issues: Our method needs supervision signals which might not always be available. Therefore, the issue of availability of data may restrict its applicability. Enforcing a causal structure on the latent representation might not be appropriate in some situations where the good latent representation factors are not structured, since training extra unnecessary layers needs more computational workload and is vulnerable to noise.

References

M. Besserve, R. Sun, and B. Schölkopf. Counterfactuals uncover the modular structure of deep generative models. *arXiv preprint arXiv:1812.03253*, 2018.

P. Brakel and Y. Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.

C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in $beta$-vae. *arXiv preprint arXiv:1804.03599*, 2018.

T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

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. *Iclr*, 2(5):6, 2017.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 517–526, 2018.

W.-N. Hsu, Y. Zhang, and J. Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pages 1878–1889, 2017.

I. Khemakhem, D. P. Kingma, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. *CoRR*, abs/1907.04809, 2019. URL http://arxiv.org/abs/1907.04809.

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

H. Kim and A. Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

J. B. Kinney and G. S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *CoRR*, abs/1709.02023, 2017. URL http://arxiv.org/abs/1709.02023.

T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015.

D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016. URL http://papers.nips.cc/book/advances-in-neural-information-processing-systems-29-2016.

F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

J. Ma, C. Zhou, P. Cui, H. Yang, and W. Zhu. Learning disentangled representations for recommendation. In *Advances in Neural Information Processing Systems*, pages 5712–5723, 2019.

E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh. Disentangling disentanglement in variational autoencoders. *arXiv preprint arXiv:1812.02833*, 2018.

I. Ng, Z. Fang, S. Zhu, and Z. Chen. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019a.

I. Ng, S. Zhu, Z. Chen, and Z. Fang. A graph autoencoder approach to causal structure learning. *CoRR*, abs/1911.07420, 2019b. URL http://arxiv.org/abs/1911.07420.

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

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

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

P. Sorrenson, C. Rother, and U. Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

R. Suter, D. Miladinović, B. Schölkopf, and S. Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. *arXiv preprint arXiv:1811.00007*, 2018.

Y. Yu, J. Chen, T. Gao, and M. Yu. Dag-gnn: Dag structure learning with graph neural networks. *arXiv preprint arXiv:1904.10098*, 2019.

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

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*, pages 9472–9483, 2018.

S. Zhu and Z. Chen. Causal discovery with reinforcement learning. *CoRR*, abs/1906.04477, 2019. URL http://arxiv.org/abs/1906.04477.

S. Zhu, I. Ng, and Z. Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations(ICLR)*, 2020.

A Proof of Proposition 1

Write the KL term in ELBO defined in Eq. 8 in the main text as

$$\begin{aligned}\mathcal{D}[q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})\|p_\theta(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})] &= \iint q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) \log \frac{q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})}{p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) p_\theta(\mathbf{z}|\mathbf{u})} d\boldsymbol{\epsilon} d\mathbf{z} \\ &= \iint q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) \log \frac{q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})}{p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})} d\boldsymbol{\epsilon} d\mathbf{z} \\ &\quad + \iint q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) \log \frac{q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})}{p_\theta(\mathbf{z}|\mathbf{u})} d\boldsymbol{\epsilon} d\mathbf{z}.\end{aligned}$$

Based on Eq. 9 in the main text, we have

$$\begin{aligned}&\iint q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) \log \frac{q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})}{p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})} d\boldsymbol{\epsilon} d\mathbf{z} \\ =&\int q_\phi(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u}) \log \frac{q_\phi(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})}{p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})} \int \delta(\mathbf{z} = \mathbf{C}\boldsymbol{\epsilon}) d\mathbf{z} d\boldsymbol{\epsilon} \\ &+ \int q_\phi(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u}) \int \delta(\mathbf{z} = \mathbf{C}\boldsymbol{\epsilon}) \log \delta(\mathbf{z} = \mathbf{C}\boldsymbol{\epsilon})) d\mathbf{z} d\boldsymbol{\epsilon} \\ =&\mathcal{D}[q_\phi(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})\|p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})] + 0 \\ =&\mathcal{D}[q_\phi(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})\|p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})],\end{aligned}$$

and

$$\begin{aligned}&\iint q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) \log \frac{q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})}{p_\theta(\mathbf{z}|\mathbf{u})} d\boldsymbol{\epsilon} d\mathbf{z} \\ =&\int q_\phi(\boldsymbol{z}|\mathbf{x}, \mathbf{u}) \log \frac{q_\phi(\boldsymbol{z}|\mathbf{x}, \mathbf{u})}{p_\theta(\mathbf{z}|\mathbf{u})} \int \delta(\boldsymbol{\epsilon} = \mathbf{C}^{-1}\boldsymbol{z}) d\boldsymbol{\epsilon} d\mathbf{z} \\ &+ \int q_\phi(\boldsymbol{z}|\mathbf{x}, \mathbf{u}) \int \delta(\boldsymbol{\epsilon} = \mathbf{C}\boldsymbol{z}) \log \delta(\boldsymbol{\epsilon} = \mathbf{C}^{-1}\boldsymbol{z}) d\boldsymbol{\epsilon} d\mathbf{z} \\ =&\mathcal{D}[q_\phi(\boldsymbol{z}|\mathbf{x}, \mathbf{u})\|p_\theta(\mathbf{z}|\mathbf{u})] + 0 \\ =&\mathcal{D}[q_\phi(\boldsymbol{z}|\mathbf{x}, \mathbf{u})\|p_\theta(\mathbf{z}|\mathbf{u})].\end{aligned}$$

Adding up the above two terms leads to the desired form of Proposition 1.

B Identifiability

B.1 Proof of Theorem 1

The general logic of the proofing follows Khemakhem et al. [2019], but we focus on both encoder and decoder. In our setting, we has joint latent variables $\boldsymbol{\epsilon}, \mathbf{z}$, and we prove identidfiabilty of both of them.

Another different setting from iVAE is that we consider a slighter transformation matrix, since our additional observations $\mathbf{u}$ of each concepts align to each causal representations $\mathbf{z}$.

Sketch of proof:

We analyze the identifiability of $\boldsymbol{\epsilon}$ starting with $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u})$. Then we define a new invertible matrix $\mathbf{L}$ which contains additional observation u_i in causal system, and use it to prove that the learned $\tilde{\mathbf{T}}$ is the transformation of $\mathbf{T}$. Step 2: We take the inference model into consideration and analyze the identifiablity of the inference model by relating the inference model to the generative model.

Details:

At the begining of proof, we consider a simple condition that the dimension of observation data d equals to the dimension of latent variables n.

The distribution has two sufficient statistics, the mean and variance of $\mathbf{z}$, which are denoted by sufficient statistics $\mathbf{T}(\mathbf{z}) = (\boldsymbol{\mu}(\mathbf{z}), \boldsymbol{\sigma}(\mathbf{z})) = (T_{1,1}(z_1), \ldots, T_{n,2}(z_n))$. We use these notations for model identifiability analysis in Section 5.

$$
\begin{aligned}
& p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u}), \\
\Rightarrow & \iint_{\mathbf{z},\boldsymbol{\epsilon}} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z},\boldsymbol{\epsilon}) p_{\boldsymbol{\theta}}(\mathbf{z},\boldsymbol{\epsilon}|\mathbf{u}) d\mathbf{z} d\boldsymbol{\epsilon} = \iint_{\mathbf{z},\boldsymbol{\epsilon}} p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{z},\boldsymbol{\epsilon}) p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{z},\boldsymbol{\epsilon}|\mathbf{u}) d\mathbf{z} d\boldsymbol{\epsilon}, \\
\Rightarrow & \int_{\mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) d\mathbf{z} = \iint_{\mathbf{z}} p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{z}) p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{z}|\mathbf{u}) d\mathbf{z}, \\
\Rightarrow & \int_{\mathbf{x}'} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{f}^{-\mathbf{1}}(\mathbf{x}')) p_{\boldsymbol{\theta}}(\mathbf{f}^{-\mathbf{1}}(\mathbf{x}')|\mathbf{u}) |\det(J_{\mathbf{f}^{-1}}(\mathbf{x}'))| d\mathbf{x}' \\
& = \int_{\mathbf{x}'} p_{\boldsymbol{\theta}}(\mathbf{x}|\tilde{\mathbf{f}}^{-\mathbf{1}}(\mathbf{x}')) p_{\widetilde{\boldsymbol{\theta}}}(\tilde{\mathbf{f}}^{-\mathbf{1}}(\mathbf{x}')|\mathbf{u}) |\det(J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{x}'))| d\mathbf{x}'.
\end{aligned} \tag{16}
$$

In determining function $\mathbf{f}$, there exist a Gaussian distribution $p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ which has infinitesimal variance. Then, the $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{f}^{-\mathbf{1}}(\mathbf{x}'))$ can be written as $p_{\boldsymbol{\xi}}(\mathbf{x} - \mathbf{x}')$. As the assumption (1) holds, this term is vanished. Then in our method, there exists the following equation:

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\mathbf{f}^{-\mathbf{1}}(\mathbf{x}')|\mathbf{u}) |\det(J_{\mathbf{f}^{-1}}(\mathbf{x}'))| &= p_{\widetilde{\boldsymbol{\theta}}}(\tilde{\mathbf{f}}^{-\mathbf{1}}(\mathbf{x}')|\mathbf{u}) |\det(J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{x}'))|, \\
\Rightarrow \widetilde{p}_{\boldsymbol{\theta}}(\mathbf{x}) &= \widetilde{p}_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}).
\end{aligned} \tag{17}
$$

Adopting the definition of multivariate Gaussian distribution, we define

$$
\boldsymbol{\lambda}_s(\mathbf{u}) = \begin{bmatrix} \lambda_1^s(u_1) & & \\ & \ddots & \\ & & \lambda_n^s(u_n) \end{bmatrix}. \tag{18}
$$

There exists the following equations:

$$
\begin{aligned}
& \log |\det(J_{\mathbf{f}^{-1}}(\mathbf{x}))| - \log \mathbf{Q}(\mathbf{f}^{-1}(\mathbf{x})) + \log \mathbf{Z}(\mathbf{u}) + \sum_{s=1}^{2} \mathbf{T}_s(\mathbf{f}^{-1}(\mathbf{x})) \boldsymbol{\lambda}_s(\mathbf{u}), \\
= & \log |\det(J_{\widetilde{\mathbf{h}}}(\mathbf{x}))| - \log \widetilde{\mathbf{Q}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \log \tilde{\mathbf{Z}}(\mathbf{u}) + \sum_{s=1}^{2} \widetilde{\mathbf{T}}_s(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) \widetilde{\boldsymbol{\lambda}}_s(\mathbf{u}),
\end{aligned} \tag{19}
$$

where $\mathbf{Q}$ denotes the base measure. In Gaussian distribution, it is $\boldsymbol{\sigma}(\mathbf{z})$.

In learning process, $\widetilde{\mathbf{A}}$ is restricted as DAG. Thus, the $\widetilde{\mathbf{C}}$ exists which is full rank matrix. The item which is not related to u in Eq. 19 are cancelled out Sorrenson et al. [2020].

$$
\sum_{s=1}^{2} \mathbf{T}_s(\mathbf{f}^{-1}(\mathbf{x})) \boldsymbol{\lambda}_s(\mathbf{u}) = \sum_{s=1}^{2} \widetilde{\mathbf{T}}_s(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) \widetilde{\boldsymbol{\lambda}}_s(\mathbf{u}) + \mathbf{b}, \tag{20}
$$

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

In our model, there exist a deterministic relationship $\mathbf{C}$ between $\boldsymbol{\epsilon}$ and $\mathbf{z}$ where $\mathbf{C} = (\mathbf{I} - \mathbf{A}^T)^{-1}$. Thus we could get equivalent of Eq. 20 as follows,

$$
\sum_{s=1}^{2} \mathbf{T}_s(\mathbf{C}\mathbf{h}(\mathbf{x})) \boldsymbol{\lambda}_s(\mathbf{u}) = \sum_{s=1}^{2} \widetilde{\mathbf{T}}_s(\widetilde{\mathbf{C}}\widetilde{\mathbf{h}}(\mathbf{x})) \widetilde{\boldsymbol{\lambda}}_s(\mathbf{u}) + \mathbf{b}`, \tag{21}
$$

where s denote the index of sufficient statistics of Gaussian distributions, indexing the mean (1) and the variance (2).

By assuming that the additional observation u_i is different, it is guaranteed that coefficients of the observations for different concepts are distinct. Thus, there exists an invertible matrix corresponding to additional information $\mathbf{u}$:

$$
\mathbf{L} = \begin{bmatrix} \boldsymbol{\lambda}_1(\mathbf{u}) & \\ & \boldsymbol{\lambda}_2(\mathbf{u}) \end{bmatrix}. \tag{22}
$$

Since the assumption that $u_i \neq 0$ holds, $\mathbf{L}$ is $2n \times 2n$ invertible and full rank diagonal matrix. Then, function of $\boldsymbol{\lambda}$ in Eq. 20 and Eq. 21 are replcaed by Eq. 22, we could get:

$$\mathbf{LT}(\mathbf{f}^{-1}(\mathbf{x})) = \widetilde{\mathbf{L}}\widetilde{\mathbf{T}}(\widetilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{b}, \tag{23}$$

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \mathbf{B_2}\widetilde{\mathbf{T}}(\widetilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{b_2}, \tag{24}$$

where

$$\mathbf{B_2} = \begin{bmatrix} \lambda_{1,1}(u_1)^{-1}\widetilde{\lambda}_{1,1}(u_1) & & \\ & \ddots & \\ & & \lambda_{n,2}(u_n)\widetilde{\lambda}_{n,2}(u_n) \end{bmatrix}. \tag{25}$$

We replace $\mathbf{f}^{-1}$ with $\mathbf{Ch}$ and we could get the equations as below:

$$\mathbf{B_3LT}(\mathbf{Ch}(\mathbf{x})) = \widetilde{\mathbf{L}}\widetilde{\mathbf{T}}(\tilde{\mathbf{C}}\tilde{\mathbf{h}}(\mathbf{x})) \Rightarrow \mathbf{T}(\mathbf{h}(\mathbf{x})) = \mathbf{B_1}\widetilde{\mathbf{T}}(\widetilde{\mathbf{h}}(\mathbf{x})) + \mathbf{b_1}, \tag{26}$$

where $\mathbf{B_3}$ is invertible matrix which corresponds to $\mathbf{C}$ and $\mathbf{B_1} = \mathbf{L^{-1}B_3^{-1}}\widetilde{\mathbf{L}}$. The definition of $\widetilde{\mathbf{L}}$ on learning model migrates the definition of $\mathbf{L}$ on ground truth.

Then we adopt the definitions following Khemakhem et al. [2019]. According to the Lemma 3 in Khemakhem et al. [2019], we are able to pick out a pair $(\epsilon_i, \epsilon_i^2)$ such that, $(\mathbf{T}_i'(z_i), \mathbf{T}_i'(z_i^2))$ are linearly independent. Then concat the two points into a vector, and denote the Jacobian matrix $\mathbf{Q} = [J_{\mathbf{T}}(\boldsymbol{\epsilon}), J_{\mathbf{T}}(\boldsymbol{\epsilon}^2)]$, and define $\tilde{\mathbf{Q}}$ on $\tilde{\mathbf{T}}(\tilde{\mathbf{h}} \circ \mathbf{Cf}(\boldsymbol{\epsilon}))$ in the same manner. By differentiating Eq. 26, we get

$$\mathbf{Q} = \mathbf{B_1}\tilde{\mathbf{Q}}. \tag{27}$$

Since the assumptiom (2) that Jacobian of $\mathbf{f}^{-1}$ is full rank holds, it can prove that both $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$ are invertible matrix. Thus from Eq. 27, $\mathbf{B_1}$ is invertible matrix. Using the same way as shown in Eq. 27, it can prove that $\mathbf{B_2}$ is invertible matrix.

Eq. 24 and Eq. 26 both hold. Combining the two results supports the identifiability result in CausalVAE.

B.2 Extension of Definition 1

In most of scenarios, latent variable is a low dimensional representation of the observation, since we are not interested in all the information in observations.

Therefore, we usually have $d > n$. We called it the reduction of dimension. We add auxiliary term as $\lambda(\mathbf{x}) = \{\lambda(\mathbf{u}), \lambda'\}$ In our model, Only n components of the latent variable are modulated, and its density has the form:

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \frac{\mathbf{Q}(\mathbf{z})}{\mathbf{Z}(\mathbf{u})} \exp \sum_{i}^{n} \mathbf{T_i}(z_i)\lambda_i(u_i) \tag{28}$$

and the term $e^{\sum_{n+1}^{d} \mathbf{T}(z_i)\lambda_i}$ is simply absorbed into $\mathbf{Q}(\mathbf{z})$. When we evaluate Eq. 19 by new definition (Eq. 28), the dimension of $p(\mathbf{z}|\mathbf{u})$ is n, because the remaining part is cancelled out.

Assume that $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})$ equal to $p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u})$. For all the observational pairs $(\mathbf{x}, \mathbf{u})$, let J_h denote the Jacobian matrix of the encoder function. Following the definition in Theorem 2 in i VAE Khemakhem et al. [2019], $\mathbf{B}$ will be indexed by 4 indicates (i, l, a, b), where $1 < i < d$ and $1 < l < s$ refer to the rows and $1 < a < d$ and $1 < b < s$ refer to the columns. We define a following equation:

$$\mathbf{v} = \tilde{\mathbf{C}} \circ \tilde{\mathbf{h}} \circ \mathbf{f}(\mathbf{z}). \tag{29}$$

The goal is to show that $v_i(\mathbf{z})$ is a function of only one z_j. We denote by $v_i^r := \frac{\partial v_i}{\partial z_r}$ and $v_i^{rt} := \frac{\partial^2 v_i}{\partial z_r \partial z_t}$. By differentiating Eq. 24 with respect to z_s, we could get:

$$T_{i,l}'(z_i) = \sum_{a=1}^{d}\sum_{b=1}^{s} B_{2,(i,l,a,b)}\tilde{T}_{a,b}'(v_a(\mathbf{z}))v_a^r(\mathbf{z}). \tag{30}$$

Lemma 1 (from Lemma 9 in Khemakhem et al. [2020]): Consider a distribution that follows a strongly exponential family. Its sufficient statistic $\tilde{\mathbf{T}}$ is differentiable almost surely. Then $\tilde{T}_i' \neq 0$ almost everywhere on $\mathbb{R}$ for all $1 \leq i \leq s$.

For $r > n$, $T'_{i,l}(z_i) = 0$, according to Lemma 1, $\tilde{T}'_{a,b}(v_a(\mathbf{z})) \neq 0$, since $\mathbf{B}_2$ is an invertible matrix, we can conclude that $v_a^r(\mathbf{z}) = 0$ for all $a < n$ and $r > n$. Therefore, we can conclude that each of the first n components of $\mathbf{v}$ is only a function of one different z_j. Thus, when $d > n$, we could get the same conclusion as Theorem 1.

C Implementation Details

We use one NVIDIA Tesla P40 GPU as our training and inference device.

For the implementation of CausalVAE and other baselines, we extend $\mathbf{z}$ to matrix $\mathbf{z} \in \mathbb{R}^{n \times k}$ where n is the number of concepts and k is the latent dimension of each $\mathbf{z}_i$. The corresponding prior or conditional prior distributions of CausalVAE and other baselines are also adjusted (this means that we extend the multivariate Gaussian to the matrix Gaussian).

The subdimensions k for each synthetic (pendulum, water) experiments are set to be 4, and 32 for CelebA experiments. The implementation of continuous DAG constraint $H(\mathbf{A})$ follows the code of Yu et al. [2019] [3].

C.1 Data Preprocessing

C.1.1 Sythetic Simulator

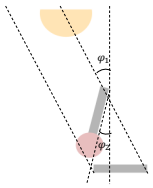


Figure 6: Generate Policy of Pendulum Simulator

Fig. 6 shows our policy of generating synthetic Pendulum data. The picture includes a pendulum. The angles of pendulum and the light are changing overtime, and projection laws are used to generate the shadows. Given the light POSITION and pendulum ANGLE, we get the angles φ_1 and φ_2. Then the system can calculate the shadow POSITION and LENGTH using triangular functions. The causal graph of concepts is shown in Fig. 7 (a). In Pendulum generator, the image size is set to be 96×96 with 4 channels. We generate about 7k images (6k for training and 1k for inference), φ_1 and φ_2 are ranged in around $[-\frac{\pi}{4}, -\frac{\pi}{4}]$.

C.1.2 Data Preprocess of CelebA

CelebA dataset contains 20K human face images. We preprocess the original dataset by following two steps:

(1) We divided the whole dataset into training dataset 85% and test dataset 15%.

(2) We only focus on facial features and resize the picture to be squared (128×128 with 3 channels).

[3] https://github.com/fishmoon1234/DAG-GNN

C.2 Intervention Experiments

C.2.1 Synthetic

In synthetic experiments, we train the model on synthetic data for 80 epochs, and use this model to generate latent code of representations. The hyperparameters of baselines are defined as default.

For CausalVAE, we set the $\alpha = 0.3$ and $(\beta, \gamma) = (1, 1)$. We use $\mathcal{N}(\mathbf{u}, |\mathbf{u}|)$ as the condition prior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$. In the implementation of CausalVAE, $|\mathbf{z}_{\text{mean}}|$ is used as the variance of condition prior.

The details of the neural networks are shown in Table 2.

C.2.2 CelebA

We also present the DO-experiments of CausalVAE and CausalGAN. In the training of the models, we use face labels (AGE, GENDER and BEARD).

For CausalVAE, we set the $\alpha = 0.3$ and $(\beta, \gamma) = (1, 1)$. We use $\mathcal{N}(\mathbf{u}, \mathbf{I})$ as the condition prior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$. For all the baseline, default hyperparameters and one common encoder and decoder structure are employed. For CausalGAN, we use the publicly available code[4].

For all the VAE-based methods, mean and variance of the distribution of the latent variable are learned during training, and the latent code z are sampled from Conditional Gaussian Distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$. In all experiments, we rescale the variance of learned representation $\mathbf{z}$ by multiplying a factor 0.1 to the original one.

Training epoches for the model is set to be 80, and our proposed CausalVAE has a pretrain step to learn causal graph $\mathbf{A}$, which takes 10 epochs.

The details of the neural networks are shown in Table 3.

C.3 The Pretrain Step for Causal Graph Learning

In our model, we need to learn the latent representation $\mathbf{z}$ and causal graph $\mathbf{A}$ simultaneously, whose optimal solution is not easy to find. Thus we adopt a pretrain stage to learn the causal graph $\mathbf{A}$ in the Mask Layer. We adopt the augmented Lagrangian to learn $\mathbf{A}$ in CausalVAE from the labels $\mathbf{u}$ in Mask Layer first. During the pretrain process, we truncate the gradient of other part of model and solve the optimization problem in Eq. 32 to learn $\mathbf{A}$.

The augmentation approach is widely used in causal discovery method, like NOTEARS Shimizu et al. [2006], DAG-GNN Yu et al. [2019]. The pretrain is a stage that learns the graph by optimizing the following objective functions:

$$\begin{aligned} &\text{minimize} \quad l_u = \mathbb{E}_{q_D} \|\mathbf{u} - \mathbf{A}^T \mathbf{u}\|_2^2 \\ &\text{subject to} \quad H(\mathbf{A}) = 0 \end{aligned} \tag{31}$$

Then, we define an augmented Lagrangian:

$$l_{pre} = l_u + \lambda H(\mathbf{A}) + \frac{c}{2} H^2(\mathbf{A}) \tag{32}$$

where λ is the Lagrangian multiplier and c is the penalty.

The following policy is used to update the λ and c:

$$\lambda_{s+1} = \lambda_s + c_s H(\mathbf{A}_s) \tag{33}$$

$$c_{s+1} = \begin{cases} c_s = \eta c_s, & if |H(\mathbf{A}_s)| > \gamma |H(\mathbf{A}_{s-1})| \\ c_s = c_s, & otherwise \end{cases}$$

where s is the iteration. In our experiments, we set $\eta = 10$ and $\gamma = \frac{1}{4}$.

[4] https://github.com/mkocaoglu/CausalGAN

encoder	decoder
4*96*96×900 fc. 1ELU	concepts×(4× 300 fc. 1ELU)
900×300 fc. 1ELU	concepts× (300×300 fc. 1ELU)
300×2*concepts*k fc.	concepts×(300× 1024 fc. 1ELU)
-	concepts×(1024× 4*96*96 fc.)

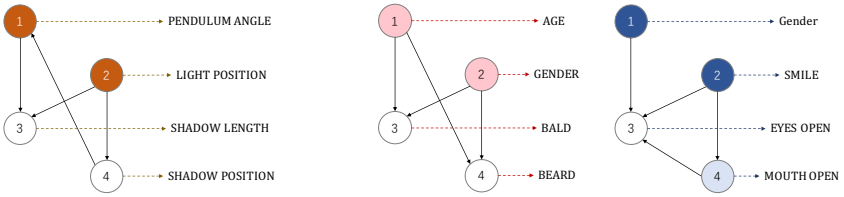
Table 2: Network design of models trained on synthetic data.

encoder	decoder
-	(1×1 conv. 128 1LReLU(0.2), stride 1)
4×4 conv. 32 1LReLU (0.2), stride 2	(4×4 convtranspose. 64 1LReLU (0.2), stride 1)
4×4 conv. 64 1LReLU (0.2), stride 2	(4×4 convtranspose. 64 1LReLU (0.2), stride 2)
4×4 conv. 64 1LReLU(0.2), stride 2	(4×4 convtranspose. 32 1LReLU (0.2), stride 2)
4×4 conv. 64 1LReLU (0.2), stride 2	(4×4 convtranspose. 32 1LReLU (0.2), stride 2)
4×4 conv. 256 1LReLU (0.2), stride 2	(4×4 convtranspose. 32 1LReLU (0.2), stride 2)
1×1 conv. 3, stride 1	(4×4 convtranspose. 3 , stride 2)

Table 3: Network design of models trained on CelebA.

D Additional Experimental Results

In this section, we show more experimental results. Fig. 7 shows the causal graph among concepts in different dataset respectively. We here show results including experiments analyzing the properties of learned representation, intervening results and the learning process of the causal graph.



(a) Pendulum (b) CelebA Group 1 (b) CelebA Group 2

Figure 7: Causal graphs of three dataset. (a) shows the causal graph in pendulum dataset. The concepts are PENDULUM ANGLE, light POSITION, SHADOW POSITION and SHADOW LENGTH. (b) shows the causal graph in CelebA, on concepts AGE, GENDER and BEARD and BALD. (c) shows the causal graph in CelebA, on concepts GENDER, SMILE, EYES OPEN and MOUTH OPEN.

D.1 The Property of Learned Representation

We test our method and baselines on both synthetic data and benchmark human face data. In the previous section, we already show the relationships between the learned representation $\tilde{\mathbf{z}}$ and the target representation $\mathbf{z}$ (related by a linear transformation formed as a diagonal matrix). In this section, we visualize it by scatter plot.

One of the important aspect of the generative model is that whether the learned representation aligns to the conditional prior we set. Our conditional prior is generated by the true label of each concept. The results show that the learned representations align to the expected representations. In figures, points are sampled from the joint distribution, and each color corresponds to one dimension.

The additional observations (labels) of Pendulum dataset and those of CelebA dataset are different. In Pendulum, the labels are values within a fixed range The labels in CelebA dataset are discrete (in $\{-1, 1\}$). Thus the scatter plots are different.

The results show that the performance of our proposed method is better than all the baselines, including the supervised method and unsupervised method.

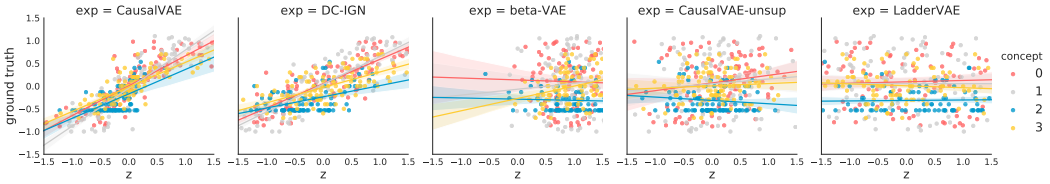


Figure 8: The figure shows the alignment of ground truth $p(\mathbf{z}|\mathbf{u})$ and the learned latent factors $q(\mathbf{z}|\mathbf{x},\mathbf{u})$ on pendulum experiments. Although DC-IGN is also the supervised method, our proposed CausalVAE shows a better performance.

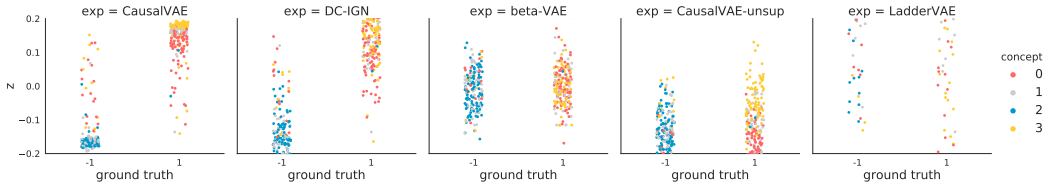


Figure 9: The figure shows the alignment of ground truth $p(\mathbf{z}|\mathbf{u})$ and the learned latent factors $q(\mathbf{z}|\mathbf{x},\mathbf{u})$ on CelebA for the concepts Group 1. The ground truth is a discrete distribution over $\{-1,1\}$, and the color of the points indicates different dimensions. The factors learned by CausalVAE show the best alignment among all.

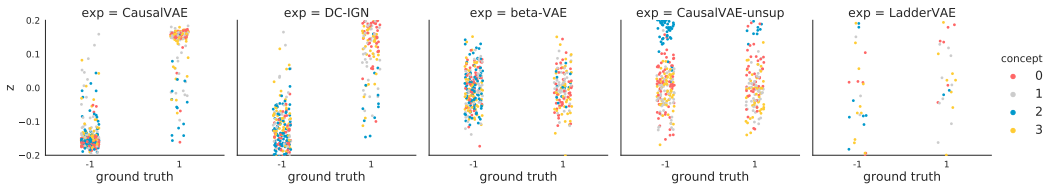
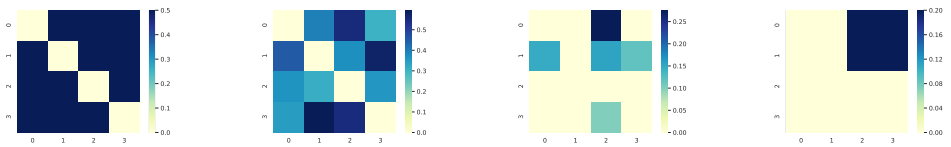


Figure 10: The figure shows the alignment between ground truth $p(\mathbf{z}|\mathbf{u})$ and the learned latent factors $q(\mathbf{z}|\mathbf{x},\mathbf{u})$ on CelebA for 5 methods (CausalVAE, DC-IGN, β-VAE, CausalVAE-unsup, LadderVAE from left to right). The ground truth is a distribution with mean taken from $\{-1,1\}$, and the color of the points indicates different dimensions. The factors learned by CausalVAE show the best alignment among all. The concepts include: 1 GENDER; 2 SMILE; 3 EYES OPEN; 4 MOUTH OPEN.

D.2 The Learned Graph



(a) Initialize (b) After 1 epoch (c) After 5 epochs (d) The true

Figure 11: Learning process of causal graph $\mathbf{A}$ in CelebA Group 1. The concepts include: 1 age; 2 GENDER; 3 BALD; 4 BEARD.

We demonstrate the learning process of causal graph in this section. Fig. 11 shows the graph learned process of CelebA Group 1. In this process, we initialize all the entries in $\mathbf{A}$ as 0.5. After 5 epochs, the graph converges. We observe an almost correct graph in this group of concepts.

D.3 Intervention Results

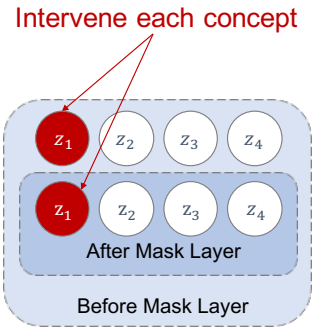


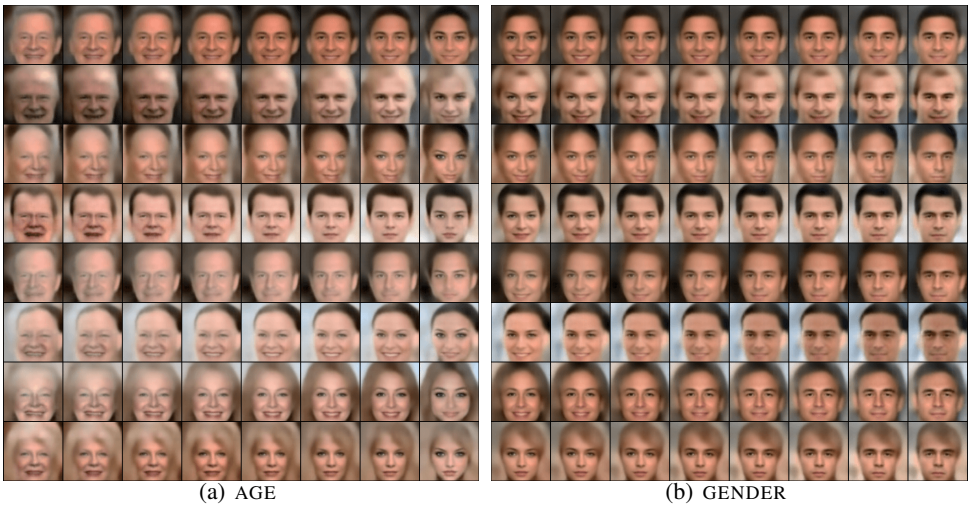
Figure 12: Intervention method

The intervention operations are as:

- For the learned model, we first put an random observed image $\mathbf{x}$ into the encoder. In this process we could get $\boldsymbol{\epsilon}$ and $\mathbf{z}$.
- Then for i-th concept, we fix the value of z_i and $g_i(\mathbf{A_i} \circ \mathbf{z})$ as constants.
- Finally, we put the new $\mathbf{z}$ into the decoder and get $\mathbf{x}'$.

The Fig. 13 demonstrates the result of CausalVAE on real world banchmark dataset CelebA Group 1, with subfigures (a) (b) (c) (d) showing the intervention experiments on concepts of AGE, GENDER, BALD and BEARD respectively. The interventions perform well that when we intervened the cause concept GENDER, the BEARD changes correspondingly. Similarly, when the cause concept AGE in intervened, its child concept BALD also changes. In contrast, intervening effect concept BEARD does not influence the causal concepts GENDER and other unrelated concepts in Fig. 13 (d). Fig. 14 demonstrates the results of CausalGAN, with subfigures (a) (b) (c) (d) showing the intervention experiments on concepts CelebA Group 1. We observe that when we intervene GENDER, the BEARD are changed. But when we intervene BEARD, concept GENDER is also changed in third line as shown by Fig. 14 (d). In general, the 'do-intervention' of CausalGAN performs worse than CausalVAE.

The Fig. 15 demonstrates the result of CausalVAE on real world banchmark dataset CelebA Group 2, with subfigures (a) (b) (c) (d) showing the intervention experiments on concepts of GENDER, SMILE, MOUTH OPEN and EYES OPEN respectively. The interventions perform well that when we intervened the cause concept GENDER, not only the appearance of GENDER but the eyes changed. When we intervened the cause concept SMILE, not only the appearance of SMILE but the MOUTH OPEN. In contrast, intervening effect concept MOUTH OPEN does not influence the causal concepts SMILE in Fig. 15 (d). Fig. 16 demonstrates the results of CausalGAN, with subfigures (a) (b) (c) (d) showing the intervention experiments on concepts CelebA Group 2. We find that when we control SMILE, the mouth is changed, as shown in the second line of Fig. 16 (b). But we find sometimes the control of SMILE influence other unrelated concepts like GENDER (shown in first line of Fig. 16 (b)). In this concepts group, CausalGAN also shows relatively unstable intervention experiments compared to that of ours.



(a) AGE (b) GENDER

(c) BALD (d) BEARD

Figure 13: Results of CausalVAE model on CelebA Group 1. The captions of the subfigures describe the controlled factors. From left to right, the pictures are results obtained by varying the value of the controlled factors.

(a) AGE (b) GENDER

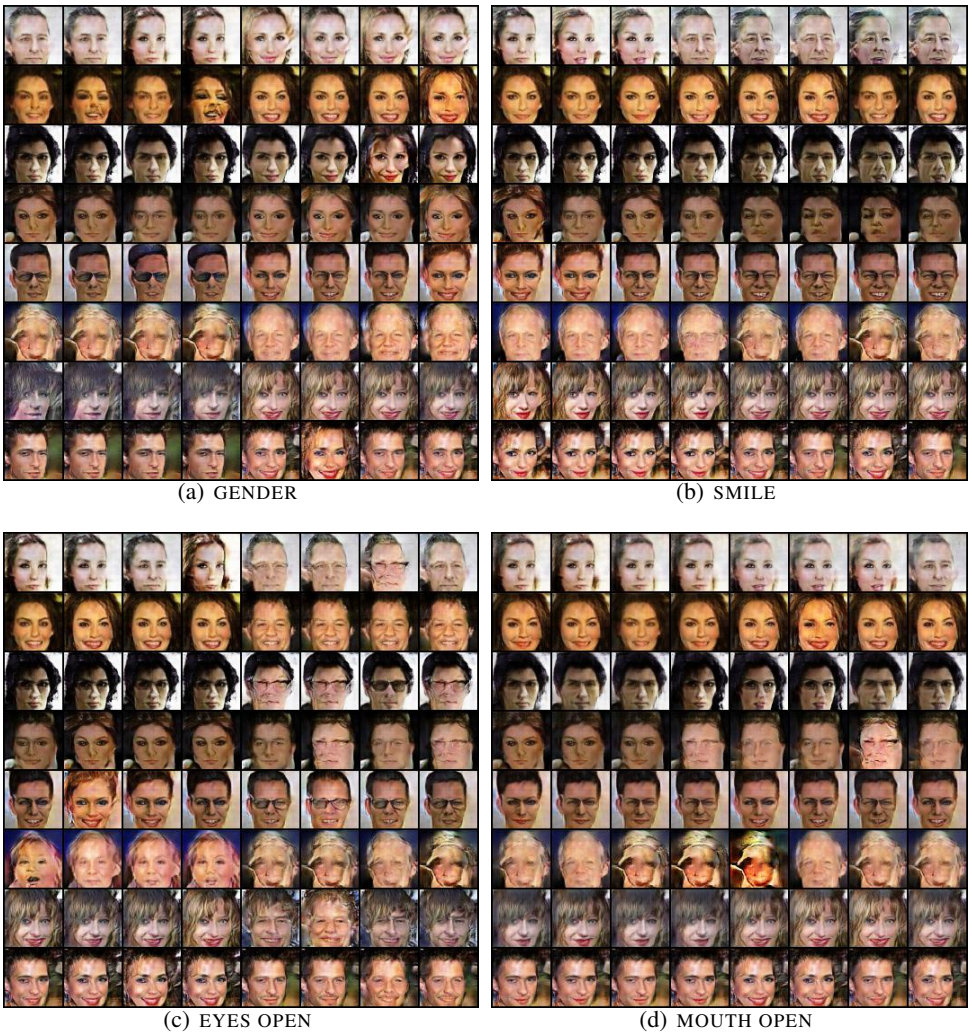
(c) BALD (d) BEARD

Figure 14: Results of CausalGAN Kocaoglu et al. [2017] model on CelebA Group 1. The captions of the subfigures describe the controlled factors. From left to right, the pictures are results obtained by varying the value of the controlled factors.

(a) GENDER (b) SMILE

(c) EYES OPEN (d) MOUTH OPEN

Figure 15: Results of CausalVAE model on CelebA Group 2. The captions of the subfigures describe the controlled factors. From left to right, the pictures are results obtained by varying the value of the controlled factors.



(a) GENDER (b) SMILE

(c) EYES OPEN (d) MOUTH OPEN

Figure 16: Results of CausalGAN model on CelebA Group 2. The captions of the subfigures describe the controlled factors. From left to right, the pictures are results obtained by varying the value of the controlled factors.