ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230258

Causally Disentangled Generative Variational AutoEncoder

SeungHwan AN[a], Kyungwoo Song[b] and Jong-June Jeon[a;*]

[a]Department of Statistics, University of Seoul, S. Korea
[b]Department of Applied Statistics, Department of Statistics and Data Science, Yonsei University, S. Korea

Abstract. We present a new supervised learning technique for the Variational AutoEncoder (VAE) that allows it to learn a causally disentangled representation and generate causally disentangled outcomes simultaneously. We call this approach Causally Disentangled Generation (CDG). CDG is a generative model that accurately decodes an output based on a causally disentangled representation. Our research demonstrates that adding supervised regularization to the encoder alone is insufficient for achieving a generative model with CDG, even for a simple task. Therefore, we explore the necessary and sufficient conditions for achieving CDG within a specific model. Additionally, we introduce a universal metric for evaluating the causal disentanglement of a generative model. Empirical results from both image and tabular datasets support our findings.

1 Introduction

Learning disentangled representation is a widely studied and challenging topic of VAE [14], and GAN [8] due to its potential to enable interpretable data generation and enhance downstream task performances [32]. Roughly speaking, studies on disentangled representation investigate the structure of a latent space where each dimension corresponds to a ground-truth factor that generates a dataset [3]. In early studies of disentangled representation, the ground-truth factors consisting of the latent space are assumed to be mutually independent exogenous variables [11]. In light of the growing interest in interpretable generative models, recent research has expanded the modeling of disentangled representation by incorporating a Structural Causal Model (SCM) and a jointly dependent model for the ground-truth factors [4, 33, 31, 25].

The importance of supervised learning for disentangled representation is raised by [19], who proved that unsupervised disentanglement learning is impossible. Especially when the model includes endogenous factors of interest, the validity of the maximum likelihood method is guaranteed only when the ground-truth factors are correctly specified [33]. So, the SCM and supervised encoder regularization method [21] are crucial for constructing latent space that causally aligns with the ground-truth factors [39, 28]. Furthermore, the alignment of the latent structure is also adapted by topological generation, which produces causally-aware generative models [37, 35, 36].

However, we found that the supervised regularization of the encoder [39, 28] is insufficient for achieving the causally-aware generative model. Even when the encoder builds a causally disentangled latent space, the causality between a latent variable and the generated output may not hold due to the entangled structure of the decoder. Our research demonstrates that causally-aware generative models are necessarily able to recover the output according to the causally disentangled factors identified by the encoder. We refer to this property as Causally Disentangled Generation (CDG) and focus on the required conditions of the decoder and the causal effect of CDG. Based on the conditions, we propose a new VAE model satisfying CDG, the CDG-VAE.

The development of CDG-VAE makes three contributions to the field of causally-aware generative models. First, we establish sufficient and necessary conditions of the decoder structure for CDG. Second, CDG-VAE can be applied to chain graphs (i.e., Partial DAGs) unlikely [37, 35, 36] that require a completely identified directed acyclic graph (DAG) for a topological generation. Third, we propose a generalized metric measuring the degree of causally disentangled generativeness under an arbitrary DAG structure of the ground-truth factors. Our metric is derived from the necessary conditions for CDG and *do*-calculus of causal effects [22].

We aim to demonstrate the effectiveness of our proposed model by evaluating two distinct types of datasets, namely image and tabular data. Specifically, we show that our model can produce causally plausible counterfactual samples with both qualitative and quantitative assessments of the image dataset. Additionally, we provide evidence of the advantages attained from the causally disentangled representation of our model in terms of two downstream tasks: sample efficiency and distributional robustness [28]. Moving on to the tabular dataset, we show that our model can generate high-quality synthetic data while preserving the observed causal structure represented by a chain graph.

Notation. Let $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p) \in \mathbb{R}^p$ be a p-dimensional vector and $\mathbf{a} = (\mathbf{a}_1, \cdots, \mathbf{a}_m)$ be a tuple where $\mathbf{a}_j \in \{1, \cdots, p\}$. $\mathbf{v}_\mathbf{a} \coloneqq (\mathbf{v}_{\mathbf{a}_1}, \mathbf{v}_{\mathbf{a}_2}, \cdots, \mathbf{v}_{\mathbf{a}_m})$ denotes a subvector of $\mathbf{v}$ sliced by $\mathbf{a}$. Let $\sigma(1), \sigma(2), \cdots, \sigma(K)$ form a partition of $\{1, 2, \cdots, p\}$ and assume that each $\sigma(j)$ is the ordered tuple whose elements are increasing. For a set $I = \{i_1, i_2, \cdots, i_k\} \subseteq \{1, \cdots, K\}$, $\sigma(I) \coloneqq \sigma(i_1) \oplus \cdots \oplus \sigma(i_k)$, where $\oplus$ is concatenation of tuples and $i_1 < i_2 < \cdots < i_k$. In particular, for $I = \{1, \cdots, K\}$ we denote $\sigma(I)$ simply by σ and we call σ the permutation inducing K-block partion on $\mathbb{R}^p$.

2 Assumptions and Model Structure

2.1 *Data Generating Process*

Let $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_p) \in \mathcal{X} \subset \mathbb{R}^p$, $\mathbf{g} = (\mathbf{g}_1, \cdots, \mathbf{g}_d) \in \mathbb{R}^d$, $\mathbf{u} = (\mathbf{u}_1, \cdots, \mathbf{u}_d) \in [0, 1]^d$ with $d < p$ be the observation, the ground-truth factor generating $\mathbf{x}$, and the annotation vector of $\mathbf{x}$. The generation and causal structure in $\mathbf{x}$ entirely depends on $\mathbf{g}$. It is

[*] Corresponding Author. Email: jj.jeon@uos.ac.kr.

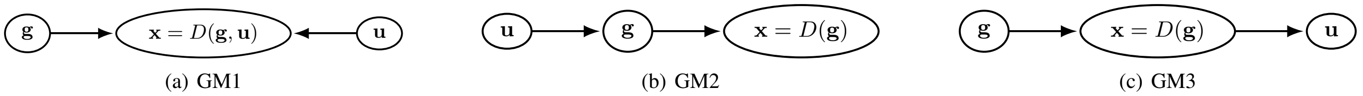


Figure 1: Various generative model (GM) structures.

assumed that $\mathbf{g}$ is unobservable, and the causality of $\mathbf{g}$ is identified only through $\mathbf{u}$.

Figure 1 shows three popular generative models (GM1, GM2, GM3) for causal disentanglement learning. In GM1 [13, 42, 10], $\mathbf{g}_i, i = 1, \cdots, d$ are mutually independent since the annotation vector $\mathbf{u}$ is directly used to generate the observation $\mathbf{x}$. In GM2 [39], $\mathbf{u}$ determines the conditional distribution of the ground-truth factor. The proposed model follows GM3 employed by [31, 28, 25]. GM3 differs from GM2 in that $\mathbf{u}$ is not explicitly used to construct the latent space. Rather, $\mathbf{u}$ is a recognized property of $\mathbf{x}$, the annotation. [21] proposes the supervised regularization ensuring that the learned latent representation is disentangled.

Assumption 1 (Blocked representation)**.** *Partition tuples $\sigma(j), \pi(j), j = 1, \cdots, K$ are given such that there are (block) causal relationships, $\mathbf{g}_{\pi(j)} \to \mathbf{x}_{\sigma(j)}, j = 1, \cdots, K$.*

[17] shows that Assumption 1 is required to infer the relationship among objects in the image. We denote the permutations inducing the K-block partition for $\mathbf{x}$ and $\mathbf{g}$ by σ and π and let σ and π be K-block consecutive partitions without loss of generality.

2.2 Embedding Causality

We introduce the Markovian SCM of $\mathbf{g}$ [22, 24], denoted by $\mathcal{M}(\mathbf{g}, \mathbf{e}, \mathbf{F}, P_{\mathbf{e}})$, where $\mathbf{g}$ and $\mathbf{e} \in \mathbb{R}^d$ are the endogenous and exogenous variables, $\mathbf{F}$ is a set of structural equations, and $P_{\mathbf{e}}$ is a probability measure of $\mathbf{e}$.

Assumption 2.
1. [Identifiable SCM] $\mathcal{M}(\mathbf{g}, \mathbf{e}, \mathbf{F}, P_{\mathbf{e}})$ is identifiable with the directed acyclic graph (DAG) $\mathcal{G}$ derived from $\mathbf{F}$.
2. [Unconfoundedness] The observation $\mathbf{x}$ is generated based on $\mathcal{M}(\mathbf{g}, \mathbf{e}, \mathbf{F}, P_{\mathbf{e}})$ and $\mathbf{x}_{\sigma(j)}$ depends on only $\mathbf{g}_{\pi(j)}$, for $j = 1, \cdots, K$.

It is ideal to employ $\mathcal{M}(\mathbf{g}, \mathbf{e}, \mathbf{F}, P_{\mathbf{e}})$ as latent variables of CDG-VAE. However, we reduce $\mathcal{M}(\mathbf{g}, \mathbf{e}, \mathbf{F}, P_{\mathbf{e}})$ to a simple non-linear structure equation model because $P_{\mathbf{e}}$ is unknown and $\mathbf{F}$ leads to heavy computational expense in training our generative model. Let $B \in \{0, 1\}^{d \times d}$ be a binary adjacency matrix whose element indicates the existence of directed edges of $\mathcal{G}$ (i.e., the causal relationships between K block subvectors of $\mathbf{g}_\pi$). For a subvector with an index s, the set of its non-descendants is denoted as $ND(s)$, and the set of its descendants is denoted as $Des(s)$.

The reduced non-linear structural equation model [40, 39, 28] is

$$f^{-1}(\mathbf{z}) = B^\top f^{-1}(\mathbf{z}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^d$ is an induced latent variable by an element-wise invertible function f and the d-dimensional standard normal distribution random variable $\boldsymbol{\epsilon}$. The distribution of latent variables in CDG-VAE is defined based on the entailed distribution of 1 and the distribution of $\boldsymbol{\epsilon}$, $p(\boldsymbol{\epsilon})$. The modeling of $p(\mathbf{z})$ through B is distinguished from conventional VAE models whose prior distributions are given without considering causality. Here, B is treated as a known matrix.

2.3 Derivation of CDG-VAE

The decoder of CDG-VAE is given by $p(\mathbf{x}_\sigma | \mathbf{z}; \theta, \beta) = \mathcal{N}(\mathbf{x}_\sigma | D(\mathbf{z}; \theta)_\sigma, \beta \cdot I)$, where $D : \mathbb{R}^d \mapsto \mathbb{R}^p$ is a function parameterized with θ, I is the $p \times p$ identity matrix, and $\beta > 0$ is the non-trainable observation noise. The rearranged vector of $D(\mathbf{z}; \theta)$ with σ is denoted as $D(\mathbf{z}; \theta)_\sigma$, where $D(\mathbf{z}; \theta)_\sigma = (D(\mathbf{z}; \theta)_{\sigma(1)}, \cdots, D(\mathbf{z}; \theta)_{\sigma(K)})$.

The proposal distribution $q(\boldsymbol{\epsilon} | \mathbf{x}; \phi)$ of CDG-VAE is given by $\mathcal{N}\big(\boldsymbol{\epsilon} | \mu(\mathbf{x}; \phi), diag(\sigma^2(\mathbf{x}; \phi))\big)$, where $\mu : \mathbb{R}^p \mapsto \mathbb{R}^d$, $\sigma^2 : \mathbb{R}^p \mapsto \mathbb{R}^d_+$ are neural networks parameterized with ϕ, and $diag(a), a \in \mathbb{R}^d$ denotes a diagonal matrix with diagonal elements a. Based on the proposal distribution, the negative ELBO (Evidence of Lower BOund) is written as

$$\begin{aligned} \mathcal{L}(\mathbf{x}; \theta, \phi, f) \quad &:= \quad \mathbb{E}_q \left[\frac{1}{2} \parallel \mathbf{x}_\sigma - D(F(\boldsymbol{\epsilon}; f, B); \theta)_\sigma \parallel^2 \right] \\ &+ \quad \beta \cdot \mathcal{KL}(q(\boldsymbol{\epsilon} | \mathbf{x}; \phi) \| p(\boldsymbol{\epsilon})) \end{aligned} \tag{2}$$

where q is the proposal distribution, $F(\boldsymbol{\epsilon}; f, B) := f((I - B^\top)^{-1} \boldsymbol{\epsilon})$, f is parameterized with the flow-based model [27, 2], $\mathcal{KL}(q \| p)$ denotes Kullback-Leibler divergence from p to q, and constant terms are omitted (see [43] Appendix A.1 for detailed derivation). See [43] Appendix A.1 for detailed covariance structures of $\mathbf{z}$. The rearranged vector of $\mathbf{z}$ with π is denoted as $\mathbf{z}_\pi$.

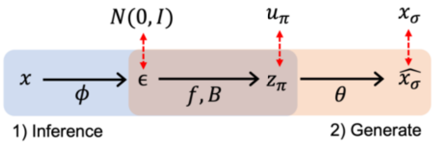


Figure 2: Model structure of CDG-VAE.

2.4 Regularization for Disentangled Representation

In the supervision setting [21] of GM3, we adopt Assumption 3, which was adopted in semi-supervised causal disentanglement learning [28]. It implies the generative model with the type of GM3, the conditional independence $\mathbf{u} \perp\!\!\!\perp \mathbf{g} | \mathbf{x}$, because $\mathbf{g}$ is $\mathcal{F}(\mathbf{x})$-measurable.

Assumption 3 (Informative supervision)**.** *$\mathbf{g}_i = \mathbb{E}[\mathbf{u}_i | \mathbf{x}]$, $\forall i$.*

Since the ground-truth factors are causally related, the representation of our encoder is entangled as the ground-truth factors [33]. With $F(\mu(\mathbf{x}; \phi); f, B)$, we aim to train the encoder parameters (ϕ, f) to approximate the SCM of $\mathbb{E}[\mathbf{u} | \mathbf{x}]$ by (1). The following definition formalizes this concept.

Definition 4 (Disentangled Representation [28])**.** *For $i = 1, \cdots, d$, suppose that there exists a 1-to-1 function h_i such that*

$$F(\mu(\mathbf{x}; \phi); f, B)_i = h_i^{-1}(\mathbb{E}[\mathbf{u}_i | \mathbf{x}]),$$

where (ϕ, f) are parameters of an encoder of a generative model, and the subscript i denotes the ith element. Then the generative model is said to have a disentangled representation with respect to $\mathbb{E}[\mathbf{u} | \mathbf{x}]$.

Note that Definition 4 implies that the disentangled representation only depends on the encoder of VAE. We adopt the supervised loss to obtain the disentangled representation, which aligns the representation and annotation vector [21]. Our final objective is minimizing

$$\mathbb{E}_{\mathbf{x}}[\mathcal{L}(\mathbf{x};\theta,\phi,f)] + \lambda \cdot \mathbb{E}_{\mathbf{x},\mathbf{u}}[\ell(F(\mu(\mathbf{x};\phi);f,B),\mathbf{u})] \tag{3}$$

with respect to (θ, ϕ, f), where $\lambda > 0$ is the tuning parameter and $\mathbb{E}_{\mathbf{x}}$, $\mathbb{E}_{\mathbf{x},\mathbf{u}}$ indicate expectations for the marginal and joint distribution of the dataset, respectively. Note that (3) is easily extended to the semi-supervised learning model because the negative ELBO $\mathcal{L}$ and the supervised loss ℓ are decoupled. In this paper, ℓ is the cross-entropy loss with the sigmoid function h_i for $\mathbf{u} \in [0,1]^d$ [28].

3 Properties of CDG-VAE

3.1 Causally Disentangled Generation

The disentanglement in the latent space has been studied with the link of explainability of VAE [11, 5, 7, 20]. The generative power of the latent space is theoretically investigated by the dimension of an activated latent space [1], which indicates a regular condition of the encoder. However, the disentangled latent space obtained by the encoder does not directly guarantee the generation of causally plausible data. We discover that the disentangled generation in the decoder is necessary.

A simple example shows that $D(\mathbf{z};\theta)_{\sigma(ND(K))}$ determined by $\mathbf{z}_{\pi(ND(K))}$ can be affected by *do*-intervention on $\mathbf{z}_i, i \in \pi(K)$ due to entangled decoder structure (see examples in Section 5.2 and [43] Appendix A.7). However, in the disentangled latent space, a total causal effect does not exist from $\mathbf{z}_i, i \in \pi(K)$ to $\mathbf{z}_j, j \in \pi(ND(K))$ because there is no directed path [23]. This result implies that causal generation requires two types of disentanglement in encoding and decoding processes, respectively. We propose a new definition of the causally disentangled generation.

Definition 5 (Causally Disentangled Generation (CDG)). *Suppose a generative model has a disentangled representation (Definition 4). Let $D(\cdot;\theta) : \mathbb{R}^d \mapsto \mathbb{R}^p$ be a decoder mean vector of the model where the input is denoted as $\mathbf{z} \in \mathbb{R}^d$. Then the model is causally disentangled generative if, for $s = 1, \cdots, K$, $D(\mathbf{z};\theta)_{\sigma(s)}$ is independent to $\mathbf{z}_{\pi(l)}, l \neq s$, given $\mathbf{z}_{\pi(s)}$.*

Definition 5 implies that CDG mimics the causal relationship of $\mathbf{g}_{\pi(j)} \to \mathbf{x}_{\sigma(j)}$ as $\mathbf{z}_{\pi(j)} \to D(\mathbf{z};\theta)_{\sigma(j)}$ for $j = 1, \cdots, K$. And the following Proposition 6 demonstrates the sufficient condition of the decoder function class satisfying CDG.

Proposition 6 (Sufficient Condition for CDG). *Suppose a generative model has a disentangled representation (Definition 4). Let $D(\cdot;\theta) : \mathbb{R}^d \mapsto \mathbb{R}^p$ be a decoder mean vector of the model. If the decoder structure of the model satisfies $D(\mathbf{z};\theta)_\sigma := \Big(D(\mathbf{z}_{\pi(1)};\theta_1), \cdots, D(\mathbf{z}_{\pi(1)};\theta_K)\Big)$, where $\theta = (\theta_1, \cdots, \theta_K)$, and $D(\cdot;\theta_j) : \mathbb{R}^{|\pi(j)|} \mapsto \mathbb{R}^{|\sigma(j)|}$ is a function parameterized with θ_j for $j = 1, \cdots, K$, then the model satisfies Definition 5.*

Proof. If a generative model satisfies Proposition 6, the jth decoder partition $D(\mathbf{z}_{\pi(j)};\theta_j)$ (which corresponds to $D(\mathbf{z};\theta)_{\sigma(j)}$) takes only a block-partitioned $\mathbf{z}_{\pi(j)}$ as input, for $j = 1, \cdots, K$. Therefore, for $j = 1, \cdots, K$, since $\mathbf{z}_{\pi(j)}$ is the direct cause of $D(\mathbf{z}_{\pi(j)};\theta_j)$, $D(\mathbf{z}_{\pi(j)};\theta_j)$ is independent to $\mathbf{z}_{\pi(i)}, i \neq j$, given $\mathbf{z}_{\pi(j)}$. □

Assumption 7 (Faithfulness). *The entailed distribution of (1) is faithful with respect to the graph induced by B.*

Proposition 8 (Existence of Total Causal Effect (TCE)). *Suppose a generative model has a disentangled representation (Definition 4) and satisfies CDG (Definition 5). Let $D(\cdot;\theta) : \mathbb{R}^d \mapsto \mathbb{R}^p$ be a decoder mean vector of the model where the input is denoted as $\mathbf{z} \in \mathbb{R}^d$. For $s = 1, \cdots, K$, under Assumption 7,*

1. there is no total causal effect from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_j$, for all $i \in \pi(s)$ and $j \in \sigma(ND(s))$.

2. if there is a directed path from $\mathbf{z}_i$ to $\mathbf{z}_j$ for some $i \in \pi(s)$ and $j \in \pi(l)$ where $l \in \{s\} \cup Des(s)$, then there is a total causal effect from $\mathbf{z}_i$ to $D(\mathbf{z};\theta)_k$, for all $k \in \sigma(l)$.

Proposition 8 states that the existence of TCE can be investigated by intervening on the latent variable if the causal structure identified by known B is embedded precisely in the latent space and a model satisfies CDG. Since TCE determines the positiveness of the causal effect, Proposition 8 motivates us to check the validity of CDG by measuring the causal effect (see Section 3.2). Due to the computational issue, we estimate the causal effect on the annotation vector instead of block partitions, based on the causal relationship from $D(\mathbf{z};\theta)_\sigma$ to $\mathbf{u}_\pi$.

First, we define the average causal effect of the latent variable on the annotation vector (the counterfactual quantity corresponding to the ground-truth factors [25]) and then propose necessary conditions for CDG based on the average causal effect. We denote $z^{(1)}, z^{(2)}$ as the vector of maximum and minimum values of latent variables given the observed dataset, respectively, and denote $\pi(s)_{-i}$ as the partition tuple $\pi(s)$ without an index i.

Definition 9 (Average Causal Effect (ACE)). *Suppose that $\mathbf{z}_i, i \in \pi(s), s = 1, \cdots, K$ is intervened with $z_i^{(1)}$ and $z_i^{(2)}$. Then, for $c = 1, \cdots, d$, the average causal effect of $\mathbf{z}_i$ on the annotation vector $\mathbf{u}_c$ given $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$ is defined as*

$$\begin{aligned}
& ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}} = z_{\pi(ND(s))\oplus\pi(s)_{-i}}) \\
:= & \Big| \mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(1)})] \\
- & \mathbb{E}[\mathbf{u}_c | z_{\pi(ND(s))\oplus\pi(s)_{-i}}, do(\mathbf{z}_i := z_i^{(2)})] \Big|.
\end{aligned}$$

Proposition 10 (Necessary Conditions for CDG). *For $i \in \pi(s), s = 1, \cdots, K$, assume that arbitrary x and $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$ are given. $z^{(j)}_{(i, z_{\pi(ND(s))}, x)}$ denotes $\mathbf{z}$ defined in (1) under intervention $do(\mathbf{z}_i := z_i^{(j)})$ given x and $z_{\pi(ND(s))\oplus\pi(s)_{-i}}$, for $j = 1, 2$. Suppose a generative model has a disentangled representation (Definition 4). For $c = 1, \cdots, d$, under Assumption 7, if the model satisfies CDG (Definition 5) and*

1. $c \in \pi(ND(s))$, then

$$ACE(\mathbf{u}_c, \mathbf{z}_i, z_{\pi(ND(s))\oplus\pi(s)_{-i}}) = 0.$$

2. there is a directed path from $\mathbf{z}_i$ to $\mathbf{z}_c$ where $c \in \pi(l)$ and $l \in \{s\} \cup Des(s)$, then

$$\begin{aligned}
0 & < ACE(\mathbf{u}_c, \mathbf{z}_i, z_{\pi(ND(s))\oplus\pi(s)_{-i}}) \\
& \leq \mathbb{E}_{p(\mathbf{x})} \Big| \mathbb{E}[\mathbf{u}_c | z^{(1)}_{(i, z_{\pi(ND(s))}, \mathbf{x})}] - \mathbb{E}[\mathbf{u}_c | z^{(2)}_{(i, z_{\pi(ND(s))}, \mathbf{x})}] \Big|,
\end{aligned}$$

where $p(\mathbf{x})$ is the probability density function of $\mathbf{x}$.

3.2 Causal Disentanglement Metric

Based on Proposition 10, we propose a metric that can evaluate how much a model is causally disentangled generative.

Definition 11 (Causal Disentanglement Metric (CDM))**.** *For $c = 1, \cdots, d$ and $i \in \pi(s), s = 1, \cdots, K$, the causal disentanglement metric (CDM) is defined as*

$$CDM(c, i) := \mathbb{E}[ACE(\mathbf{u}_c, \mathbf{z}_i, \mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}})],$$

where $\mathbb{E}$ indicates the expectation with respect to $\mathbf{z}_{\pi(ND(s))\oplus\pi(s)_{-i}}$.

First, CDM for the first case of Proposition 10 measures interventional robustness [31]. If a generative model satisfies CDG, ACE should be zero by Proposition 10 (i.e., no total causal effect), and CDM is also exactly zero. Thus, $CDM(c, i) > 0$ implies that the model is not interventional robust.

Next, CDM for the second case of Proposition 10 measures counterfactual generativeness [25]. Suppose a generative model satisfies CDG. Then, ACE should be non-zero (i.e., total causal effect exists) and is the lower bound of the coverage for the counterfactual quantity $\mathbf{u}_c$ by Proposition 10. Therefore, $CDM(c, i) = 0$ or $CDM(c, i) \approx 0$ implies that the model lacks counterfactual generativeness.

[31, 25, 28] have considered disentangled causal mechanisms and proposed metrics for causal disentanglement. Unlike our study, the existing studies do not deal with a causal effect between latent variables nor derive the metric based on the SCM of the ground-truth factors. In particular, our metric is theoretically justified by the necessary conditions for the CDG and the do-calculus of causal effects. Therefore, CDM can be regarded as a generalized causal disentanglement metric under the arbitrary DAG structure of ground-truth factors. See [43] Appendix A.4 for identification of CDM's upper and lower bounds.

Also, the faithfulness condition (Assumption 7) excludes the case where the directed path exists in B, but the total causal effect does not under. It implies that, under the faithfulness condition, our proposed metric CDM is valid to measure the causally disentangled generativeness of the model. However, without the faithfulness condition, we can only measure the interventional robustness of Proposition 10. Therefore, the faithfulness condition is required to measure the causal disentanglement of the model by the average causal effect.

4 Related Work

Active Latent Dimensions. [6, 29] propose the statistics to detect active latent dimensions, which encode useful information about the data and are significant in data generation. [1] shows that the posterior variance $\sigma^2(\mathbf{x}; \phi)$ determines whether the latent dimension affects generated data. However, the definition of disentangled representation (Definition 4) only depends on the deterministic component $\mu(\mathbf{x}; \phi)$. Therefore, a latent dimension can be simultaneously non-active and disentangled. It is a critical issue because the causally plausible counterfactual samples can not be generated when non-active latent variables are intervened (see [43] Appendix A.8 for an example). [26] mitigates the non-activation issue by maximizing the additional mutual information regularization term. However, we discover that all latent dimensions are always active under Proposition 6.

Supervised Causal Disentanglement Learning. [39, 28] propose causal disentanglement learning methods based on supervision setting [21] and embed the SCM in the latent space to make the representation causally entangled. To obtain the disentangled representation, [39] aligns the latent variable and the annotation vector using KL-divergence and the customized prior, and [28] regularizes the encoder. However, since [39, 28] only constrain the encoder, their methods can not guarantee that generating causally plausible data is achievable.

Causally-Aware Synthetic Data Generation. Since tabular datasets are already well-structured, covariates (columns) are usually assumed to be causally related [36]. To exploit causations in the synthetic data generation, [37, 36, 35] generate data in the order of causal topology, and consequently, they require the completely identified DAG, not the Partial DAG. However, from the observational data, the true causal graph of covariates can be identified only up to a Markov Equivalence Class (MEC), including undirected edges. Even though [35] proves that their generator converges to the right distribution for any graph belonging to MECs, incorrect edge directions have the potential risk of misunderstandings of causations.

5 Experiments

This section demonstrates that our model is causally disentangled in both the encoding and decoding processes. Our numerical experiments show that CDG-VAE can achieve two goals: 1) the causally plausible counterfactual generation under interventions on latent variables and 2) synthetic data generation preserving the observed causal structure. Furthermore, the performances of our model, sample efficiency, distributional robustness, and synthetic data quality, are presented with three downstream tasks. The code and appendix are available at https://github.com/an-seunghwan/CDG-VAE.

5.1 Overview

Dataset. For evaluation, we consider two types of datasets, image and tabular. For an image dataset, a simulated pendulum dataset [28, 39] is used. And `loan`, `adult`, and `covertype` datasets are used for real tabular datasets (see [43] Appendix A.6 for detailed data descriptions).

Compared Models. We train the vanilla VAE and InfoMax VAE [26] based on the objective function (3) (see [43] Appendix A.1 for detailed objective functions). We also compare existing disentangled generative models (CausalVAE [39], DEAR [28]) and synthesizers (TVAE [38], CTAB-GAN [41]).

Note that all models are trained under the ground-truth causal graph, not a super-graph, because we numerically find that DEAR and CausalVAE are not able to discover ground-truth causal relationships. VAE, InfoMax VAE, and CDG-VAE share the same network architecture for the encoder; however, only CDG-VAE has the decoder structure of Proposition 6. Notably, all models have the same size of the latent dimension.

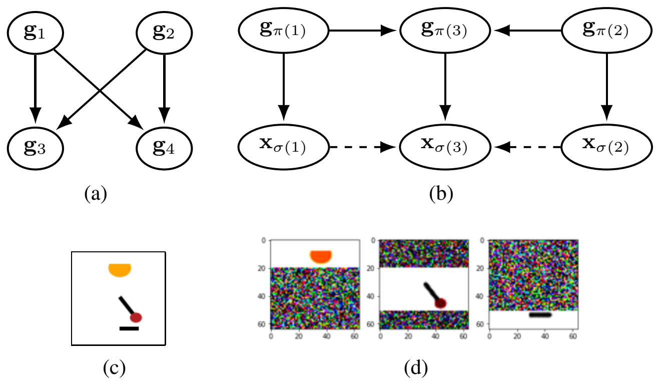
5.2 Image Dataset

In the pendulum dataset, there exist four ground-truth factors: $\mathbf{g}_1$(light angle), $\mathbf{g}_2$(pendulum angle), $\mathbf{g}_3$(shadow length), and $\mathbf{g}_4$(shadow position). These factors have the causal relationship given as the DAG structure of Figure 3(a). Partition tuples of $\mathbf{g}_\pi$ are $\pi(1) = (1), \pi(2) = (2)$, and $\pi(3) = (3, 4)$. Causal structures are visualized in Figure 3(b).

As in [28], we introduce random measurement noises in the generation of annotation vectors to make the pendulum dataset more realistic. Shadows of 20% corrupted data are randomly generated to mimic some environmental disturbance. The training and test dataset sizes are 7,500 and 2,500, respectively. We also evaluated our model under a semi-supervised setting where only 10% of the annotation vectors are available. Since the annotations are bounded from 0 to 1, CDM's upper and lower bounds are also bounded from 0 to 1.

5.2.1 Causally Disentangled Generation

We investigate the performance of counterfactual generation through *do*-intervention on the latent variable, and Figure 4 shows corresponding generated images. For CausalVAE and DEAR, if we intervene on



(a) (b)

(c) (d)

Figure 3: Pendulum dataset. (a) DAG of the ground-truth factors $\mathbf{g} = (\mathbf{g}_1, \cdots, \mathbf{g}_4)$. (b) The causal relationships between $\mathbf{g}_\pi$ and $\mathbf{x}_\sigma$. Dashed edges indicate induced edges by causations of $\mathbf{g}_\pi$. (c) An observation example. (d) From left to right, $\mathbf{x}_{\sigma(1)}$ (the light), $\mathbf{x}_{\sigma(2)}$ (the pendulum), and $\mathbf{x}_{\sigma(3)}$ (the shadow).

shadow length and position, block partitions of their parents (i.e., light angle and pendulum angle) are affected (see the third and fourth row of Figure 4(a) and 4(b)). However, CDG-VAE can generate images in which block partitions of light angle and pendulum angle are not affected when their children (i.e., shadow length and position) are intervened (see the third and fourth row of Figure 4(c)). Therefore, CDG-VAE under Proposition 6 enables CDG.

5.2.2 *Causal Disentanglement Metric*

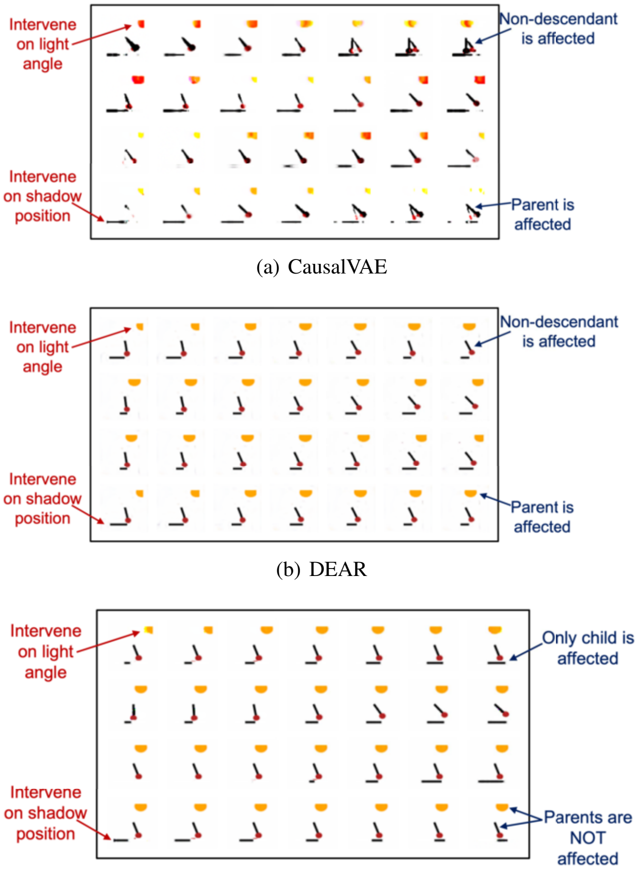
We compared models using our proposed causal disentanglement metric CDM. For ease of explanation, we use $CDM(light, length)$ instead of $CDM(c = 1, i = 3)$. The second and third columns of Table 1 indicate the interventional robustness. For example, as light angle is a parent of shadow length, $CDM(light, length)$ measures interventional robustness. CDG-VAE achieves exactly zero in both $CDM(light, length)$ and $CDM(angle, pos)$, which implies that only CDG-VAE can satisfy CDG. It is worth mentioning that CDG-VAE achieves exactly zero CDM values for all other cases (see [43] Appendix A.8).

And the fourth and fifth columns of Table 1 show the counterfactual generativeness. For example, since shadow length is a child of pendulum angle, $CDM(length, angle)$ measures the counterfactual generativeness. CDG-VAE achieves the highest score in $CDM(pos, pos)$ and outperforms CausalVAE and DEAR models in $CDM(length, angle)$. Therefore, Table 1 indicates that CDG-VAE has competitive counterfactual generativeness performance (see [43] Appendix A.8 for other cases).

5.2.3 *Downstream Task*

This section investigates the advantages of causally disentangled representations for two downstream tasks: sample efficiency and distributional robustness [28]. The binary classification task is mainly used for evaluation, and we generate the target label as a function of the ground-truth factors. It means that the representations learned from the generative model are causal representations of the target label (see [43] Appendix A.7 for a detailed explanation).

Sample Efficiency. To measure the sample efficiency, we use the statistical efficiency score defined as the test accuracy based on 100 samples divided by the test accuracy based on all samples, following [19, 28]. We use fitted encoders to extract representations and train an MLP classifier on top of the representations to predict the target label. Notably, all models are evaluated with the same MLP. We also report the test accuracies to prevent misleading when a classifier achieves poor test accuracy in both cases. Table 2 shows that CDG-VAE performs the best in the sample efficiency downstream task.



(a) CausalVAE

(b) DEAR

(c) CDG-VAE(NL)

Figure 4: Examples of generated counterfactual images. For each image, intervened dimensions are light angle, pendulum angle, shadow length, and shadow position from top to bottom. 'NL' denotes that a nonlinear f is used.

Distributional Robustness. To evaluate the distributional robustness of causal representations, we manipulate the training dataset to impose spurious correlations between the target label and some spurious features of the image. We choose $background\_color \in \{white(0), blue(1)\}$ as a spurious feature [28]. 80% of the training samples have the same value of the target label and $background\_color$ (strong correlation), but the test samples do not have such correlations (all label values are distributed equally).

Table 3 shows the performances of the compared models in the distributional robustness (downstream task) of the causally disentangled representation. The worst case ('TrainWorst' and 'TestWorst') is when the target label and the spurious feature $background\_color$ are grouped to have the opposite label. For that group, the spurious feature has a different correlation between the training and test datasets. And Table 3 shows that CDG-VAE shows the best test accuracy in the worst cases. Therefore, CDG-VAE can produce a causally disentangled representation robust to distributional shifts. Note that 'TrainAvg' or 'TrainWorst' are not reasonable criteria to judge the

Table 1: Numbers in parentheses are lower and upper bounds of CDM. ‘L’ and ‘NL’ denote the model with linear and nonlinear f, and ‘*’ denotes the semi-supervised learned model. Mean and standard deviation values are obtained from 10 repeated experiments. ‘pos’ denotes shadow position. ↑ denotes higher is better and ↓ denotes lower is better.

	Interventional Robustness ↓		Counterfactual Generativeness ↑	
Model	$CDM(light, length)$	$CDM(angle, pos)$	$CDM(length, angle)$	$CDM(pos, pos)$
VAE(L)	$(0.44, 0.44)_{\pm(0.35,0.35)}$	$(0.28, 0.28)_{\pm(0.30,0.31)}$	$(0.31, 0.32)_{\pm(0.16,0.15)}$	$(0.27, 0.28)_{\pm(0.25,0.24)}$
VAE(NL)	$(0.38, 0.40)_{\pm(0.28,0.27)}$	$(0.27, 0.33)_{\pm(0.25,0.24)}$	$(0.33, 0.34)_{\pm(0.12,0.12)}$	$(0.31, 0.34)_{\pm(0.21,0.20)}$
InfoMax(L)	$(0.42, 0.43)_{\pm(0.39,0.38)}$	$(0.38, 0.38)_{\pm(0.34,0.34)}$	$(0.40, 0.40)_{\pm(0.26,0.25)}$	$(0.29, 0.31)_{\pm(0.22,0.20)}$
InfoMax(NL)	$(0.37, 0.39)_{\pm(0.32,0.30)}$	$(0.26, 0.33)_{\pm(0.28,0.25)}$	$\mathbf{(0.44, 0.44)}_{\pm(0.21,0.21)}$	$(0.31, 0.34)_{\pm(0.19,0.16)}$
CausalVAE	$(0.28, 0.28)_{\pm(0.11,0.10)}$	$(0.17, 0.17)_{\pm(0.09,0.08)}$	$(0.10, 0.10)_{\pm(0.04,0.04)}$	$(0.29, 0.29)_{\pm(0.09,0.09)}$
DEAR	$(0.21, 0.23)_{\pm(0.16,0.15)}$	$(0.26, 0.29)_{\pm(0.25,0.24)}$	$(0.23, 0.25)_{\pm(0.23,0.23)}$	$(0.16, 0.20)_{\pm(0.18,0.16)}$
CDG-VAE(L)	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$(0.24, 0.25)_{\pm(0.10,0.09)}$	$(0.69, 0.69)_{\pm(0.25,0.25)}$
CDG-VAE(NL)	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$(0.35, 0.36)_{\pm(0.16,0.15)}$	$\mathbf{(0.78, 0.78)}_{\pm(0.24,0.24)}$
CDG-VAE(L)*	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$(0.21, 0.22)_{\pm(0.09,0.07)}$	$(0.66, 0.66)_{\pm(0.22,0.22)}$
CDG-VAE(NL)*	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$\mathbf{(0.00, 0.00)}_{\pm(0.00,0.00)}$	$(0.29, 0.30)_{\pm(0.12,0.11)}$	$\mathbf{(0.79, 0.79)}_{\pm(0.21,0.21)}$

Table 2: Sample efficiency and test accuracies with 100 and all training samples. ‘L’ and ‘NL’ denote the model with linear and nonlinear f, and ‘*’ denotes the semi-supervised learned model. Mean and standard deviation values are obtained from 10 repeated experiments. Higher is better.

Model	100(%)	All(%)	SE(100/All)
VAE(L)	$88.41_{\pm 2.07}$	$90.18_{\pm 1.01}$	$98.03_{\pm 1.67}$
VAE(NL)	$87.74_{\pm 1.23}$	$90.16_{\pm 0.60}$	$97.32_{\pm 1.25}$
InfoMax(L)	$88.69_{\pm 1.36}$	$90.34_{\pm 0.52}$	$98.18_{\pm 1.45}$
InfoMax(NL)	$87.74_{\pm 1.05}$	$90.15_{\pm 0.54}$	$97.32_{\pm 0.98}$
CausalVAE	$50.39_{\pm 2.08}$	$86.94_{\pm 1.44}$	$57.98_{\pm 2.74}$
DEAR	$82.92_{\pm 3.45}$	$88.92_{\pm 1.34}$	$93.27_{\pm 3.97}$
CDG-VAE(L)	$89.48_{\pm 0.76}$	$90.68_{\pm 0.24}$	$\mathbf{98.67}_{\pm 0.81}$
CDG-VAE(NL)	$87.94_{\pm 1.29}$	$90.19_{\pm 0.52}$	$97.51_{\pm 1.22}$
CDG-VAE(L)*	$89.33_{\pm 0.63}$	$90.52_{\pm 0.34}$	$\mathbf{98.69}_{\pm 0.48}$
CDG-VAE(NL)*	$88.39_{\pm 0.69}$	$90.06_{\pm 0.47}$	$98.14_{\pm 0.90}$

best model because Table 3 shows the distributional robustness where the distribution of the test dataset is changed (i.e., distributional shift).

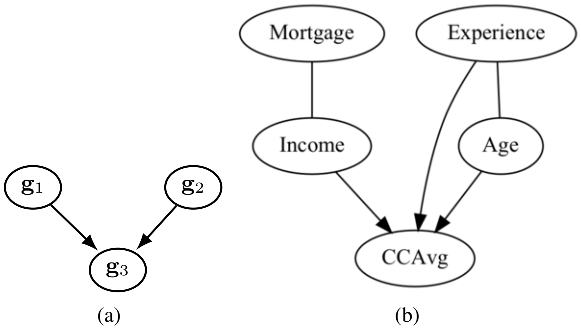
On the other hand, if the latent space is not causally disentangled, the features of observation partitions are entangled in the encoder, and the model exploits the entangled features in the learning of the latent space. It results that the latent variable is affected by the changes in the observation partitions, which are not causally related. See the toy example of such a case in [43] Appendix A.7. And we guess that the latent space of CausalVAE is not fully disentangled, and the downstream classifier utilizes the correlation information between the spurious feature and the target label. Consequently, the downstream classifier with CausalVAE is overfitted and shows a higher score in ‘TrainAvg’ metric in Table 3.

5.3 Tabular Datasets

5.3.1 Causal Disentanglement Learning

In the supervised causal disentanglement learning method with tabular datasets, we assume that the causal dependencies between K block subvectors of $\mathbf{x}_\sigma$ are induced by the causal structure of $\mathbf{g}_\pi$. Therefore, $\mathbf{x}_\sigma$ is a chain graph-structured data, such as a multi-layered proteomic data [9] and $\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma(K)}$ are chain components [16].

Figure 5(b) shows the chain graph of `loan` dataset obtained by PC algorithm [30]. The chain components of $\mathbf{x}_\sigma$ are $\mathbf{x}_{\sigma(1)} =$ (Mortgage, Income), $\mathbf{x}_{\sigma(2)} =$ (Experiences, Age), and $\mathbf{x}_{\sigma(3)} =$ (CCAvg). We assume that each chain component $\mathbf{x}_{\sigma(j)}$ is generated by the single ground-truth factor $\mathbf{g}_j$, as in GM3 of Figure 1 ($\mathbf{g}_j \rightarrow \mathbf{x}_{\sigma(j)}$). Thus, Figure 5(a) is the DAG structure of the ground-truth factors $\mathbf{g}$.



(a)

(b)

Figure 5: `loan` dataset. (a) DAG structure of ground-truth factors $\mathbf{g}$. (b) The chain graph of covariates.

Due to undirected edges between covariates (e.g., the edge between Mortgage and Income), the SCM of covariates is not defined well, and the covariate-wise topological generation of [37, 35, 36] is not applicable. However, without a completely identified DAG, CDG-VAE can achieve the CDG. That is, our model can include (Mortgage, Income) and (Experience, Age) as the chain components in disentanglement learning.

First, since Figure 5(a) consists of only directed causal relationships, the SCM of the latent variables can be formulated based on the DAG of Figure 5(a). Next, we define the bijection output of each chain component $\mathbf{x}_{\sigma(j)}$ as $\mathbb{E}[\mathbf{u}_j|\mathbf{x}_{\sigma(j)}]$ and assume that $\mathbf{g}_j = \mathbb{E}[\mathbf{u}_j|\mathbf{x}_{\sigma(j)}] = \mathbb{E}[\mathbf{u}_j|\mathbf{x}]$, for $j = 1, 2, 3$. The DAG obtained by the PC algorithm with bijection outputs is equivalent to Figure 5(a), implying that bijection outputs have the same causal structure as the ground-truth factors. In practice, we use the interleaving function for bijection after the min-max scaling (note that the independence is not affected by scaling). See [43] Appendix A.6 for other tabular datasets' chain graph structures and chain components.

5.3.2 Synthetic Data Generation

We evaluated the performance of our model in synthetic data generation. To measure whether the observed causal structure is preserved, we use the Structural Hamming Distance (SHD) [34] between causal graphs of the observed and a synthetic dataset. Causal graphs are

Table 3: Distributional robustness: Train and test dataset accuracy for average ('Avg') and worst ('Worst') cases. 'L' and 'NL' denote the model with linear and nonlinear f, and '*' denotes the semi-supervised learned model. Mean and standard deviation values are obtained from 10 repeated experiments. Higher is better.

Model	TrainAvg(%)	TrainWorst(%)	TestAvg(%)	TestWorst(%)
VAE(L)	$61.70_{\pm 2.46}$	$57.70_{\pm 2.56}$	$58.88_{\pm 0.85}$	$55.73_{\pm 2.75}$
VAE(NL)	$62.14_{\pm 1.98}$	$57.63_{\pm 2.93}$	$59.41_{\pm 0.93}$	$55.91_{\pm 3.34}$
InfoMax(L)	$61.50_{\pm 2.08}$	$57.55_{\pm 3.56}$	$59.04_{\pm 1.21}$	$56.00_{\pm 3.29}$
InfoMax(NL)	$62.33_{\pm 2.35}$	$58.07_{\pm 2.48}$	$\mathbf{59.60}_{\pm 0.77}$	$56.45_{\pm 2.36}$
CausalVAE	$\mathbf{73.93}_{\pm 0.99}$	$35.52_{\pm 5.57}$	$57.92_{\pm 1.28}$	$33.91_{\pm 4.92}$
DEAR	$62.33_{\pm 2.27}$	$55.62_{\pm 4.61}$	$58.60_{\pm 1.02}$	$53.16_{\pm 4.25}$
CDG-VAE(L)	$64.28_{\pm 5.22}$	$50.45_{\pm 10.07}$	$58.02_{\pm 0.91}$	$48.59_{\pm 10.05}$
CDG-VAE(NL)	$60.97_{\pm 0.91}$	$\mathbf{59.34}_{\pm 1.30}$	$59.22_{\pm 0.60}$	$\mathbf{57.21}_{\pm 1.50}$
CDG-VAE(L)*	$70.43_{\pm 4.69}$	$40.47_{\pm 11.16}$	$55.36_{\pm 1.37}$	$36.31_{\pm 9.87}$
CDG-VAE(NL)*	$67.19_{\pm 3.61}$	$44.28_{\pm 9.88}$	$55.66_{\pm 1.96}$	$41.01_{\pm 9.48}$

Table 4: SHD and data quality scores for synthetic datasets. All models use linear f. Mean and standard deviation values are obtained from 10 repeated experiments. 'Baseline' indicates results from the observed dataset. ↑ denotes higher is better and ↓ denotes lower is better.

Dataset	`loan`$(p = 5)$		`adult`$(p = 5)$		`covertype`$(p = 8)$	
Model	SHD ↓	R^2 ↑	SHD ↓	F_1 ↑	SHD ↓	F_1 ↑
Baseline	-	0.392	-	0.818	-	0.712
VAE	$6.1_{\pm 2.3}$	$-7.936_{\pm 15.127}$	$7.0_{\pm 2.3}$	$0.739_{\pm 0.032}$	$17.7_{\pm 3.1}$	$0.067_{\pm 0.022}$
InfoMax	$7.1_{\pm 1.1}$	$-5.149_{\pm 7.920}$	$7.5_{\pm 1.2}$	$0.712_{\pm 0.061}$	$17.8_{\pm 4.1}$	$0.102_{\pm 0.021}$
TVAE	$5.0_{\pm 1.8}$	$-0.631_{\pm 0.380}$	$5.1_{\pm 1.5}$	$0.724_{\pm 0.009}$	$16.2_{\pm 2.9}$	$\mathbf{0.358}_{\pm 0.024}$
CTAB-GAN	$4.9_{\pm 0.9}$	$-0.912_{\pm 0.504}$	$5.5_{\pm 3.2}$	$\mathbf{0.795}_{\pm 0.007}$	$17.3_{\pm 3.6}$	$0.077_{\pm 0.029}$
CDG-VAE	$0.9_{\pm 0.3}$	$-0.982_{\pm 1.663}$	$0.3_{\pm 0.9}$	$0.696_{\pm 0.003}$	$\mathbf{1.6}_{\pm 0.5}$	$0.127_{\pm 0.015}$
CDG-TVAE	$\mathbf{0.4}_{\pm 0.5}$	$\mathbf{0.013}_{\pm 0.010}$	$\mathbf{0.2}_{\pm 0.4}$	$0.645_{\pm 0.001}$	$2.8_{\pm 0.6}$	$0.178_{\pm 0.005}$

obtained by the PC algorithm [30]. The smaller SHD value indicates a model can generate synthetic data with precise causal relationship information.

On the other hand, to evaluate the synthetic data quality, we use the synthetic data as training data for three widely used machine learning algorithms: linear (logistic) regression, Random Forest, and Gradient Boosting (see [43] Appendix A.6 for details). We average the following metrics: R^2 for the regression and F_1 for the classification problem. Note that the synthetic data and the observed data have the same size. Here, CDG-TVAE is the new synthesizer where the mode-specific normalization technique of TVAE is combined with CDG-VAE.

The SHD values of Table 4 show that models under Proposition 6 (CDG-VAE and CDG-TVAE) outperform in terms of preserving the original causal structure well compared to the other models in synthetic data generation. Moreover, we observe that the difference in the SHD value between the proposed model and the compared models becomes significantly larger as the number of covariates (p) increases.

Although CDG-TVAE compromises the F_1 score and SHD value (`covertype` dataset) since we restrict the latent space and the decoder structure of the model according to the causal relationships, CDG-TVAE has a competitive performance in data quality scores in Table 4 (`loan`, `adult` datasets). Therefore, CDG-TVAE can generate synthetic data, which can be used as a good proxy of the original data while preserving the observed causal structure.

6 Conclusion and Limitations

We demonstrate that causally disentangled latent space is insufficient to generate an image according to the causal mechanism defined by a general SCM. Figure 4 shows that the entangled decoder can not generate the counterfactual image even when the causal structure is simple and the block-partition indices are known. This result concludes that the disentanglement of features is crucial both in the encoder and the decoder, and a causal relationship is contaminated without the disentanglement.

The decoder structure of Proposition 6 is very restricted since it requires the indices of all partition blocks (i.e., obtaining complete causal information for supervision). We do not believe the proposed decoder structure is directly applicable to large-scale datasets such as the CelebA dataset [18]. Besides, in synthetic data generation, we numerically find that the supervised disentanglement learning with interleaving bijection (Section 5.3) can preserve the observed causal structure only when the cardinality of a chain component is small (empirically, less than 4). We guess that it is because the interleaving function has limited expressiveness.

We expect that we can effectively deal with the partitioned blocks as a multi-layered image with the same size and control the property of disentangled features by utilizing mutual information between generated samples and annotation vectors. In addition, to enrich the expressiveness of the latent space with a causal structure, it is necessary to find computationally tractable bijections maps for multivariates and multi-layered data and apply supervised disentanglement learning. We leave solving the two problems as our future work.

Lastly, our model is identifiable in the sense of [15] (the generalized identifiability). But it is not sure that our model is also A-identifiable or P-indentifiable [12] because our standard Gaussian assumption for priors violates the sufficient condition of the identifiability. We will thoroughly discuss the identifiability of our proposed model and address this as a crucial research topic for future investigations.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022R1A4A3033874, No. NRF-2022R1F1A1074758, and NO.2022M3J6A1084845). The authors acknowledge the Urban Big data and AI Institute of the University of Seoul supercomputing resources (http://ubai.uos.ac.kr) made available for conducting the research reported in this paper.

References

[1] SeungHwan An and Jong-June Jeon, 'Customized latent space: Practical usage of variational autoencoder', *SSRN Electronic Journal*, (2023).

[2] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen, 'Invertible residual networks', in *International Conference on Machine Learning*, pp. 573–582. PMLR, (2019).

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent, 'Representation learning: A review and new perspectives', *IEEE transactions on pattern analysis and machine intelligence*, **35**(8), 1798–1828, (2013).

[4] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal, 'A meta-transfer objective for learning to disentangle causal mechanisms', *arXiv preprint arXiv:1901.10912*, (2019).

[5] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin, 'Multilevel variational autoencoder: Learning disentangled representations from grouped observations', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, (2018).

[6] Yuri Burda, Roger Baker Grosse, and Ruslan Salakhutdinov, 'Importance weighted autoencoders', *CoRR*, **abs/1509.00519**, (2015).

[7] Emilien Dupont, 'Learning disentangled joint continuous and discrete representations', *Advances in Neural Information Processing Systems*, **31**, (2018).

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial networks', *Communications of the ACM*, **63**(11), 139–144, (2020).

[9] Min Jin Ha, Francesco Claudio Stingo, and Veerabhadran Baladandayuthapani, 'Bayesian structure learning in multilayered genomic networks', *Journal of the American Statistical Association*, **116**(534), 605–618, (2021).

[10] Sina Hajimiri, Aryo Lotfi, and Mahdieh Soleymani Baghshah, 'Semi-supervised disentanglement of class-related and class-independent factors in vae', *ArXiv*, **abs/2102.00892**, (2021).

[11] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, 'beta-vae: Learning basic visual concepts with a constrained variational framework', (2016).

[12] Ilyes Khemakhem, Diederik P. Kingma, and Aapo Hyvärinen, 'Variational autoencoders and nonlinear ica: A unifying framework', in *International Conference on Artificial Intelligence and Statistics*, (2019).

[13] Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, 'Semi-supervised learning with deep generative models', in *NIPS*, (2014).

[14] Diederik P Kingma and Max Welling, 'Auto-encoding variational bayes', *arXiv preprint arXiv:1312.6114*, (2013).

[15] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam, 'Identifiability of deep generative models without auxiliary information', in *Neural Information Processing Systems*, (2022).

[16] Daphne Koller and Nir Friedman, *Probabilistic graphical models: principles and techniques*, MIT press, 2009.

[17] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien, 'Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica', in *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, (2022).

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, 'Deep learning face attributes in the wild', in *Proceedings of International Conference on Computer Vision (ICCV)*, (December 2015).

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

[20] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem, 'Disentangling factors of variation using few labels', *arXiv preprint arXiv:1905.01258*, (2019).

[21] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem, 'Disentangling factors of variations using few labels', in *International Conference on Learning Representations*, (2020).

[22] Judea Pearl, *Causality*, Cambridge university press, 2009.

[23] J. Peters, Dominik Janzing, and Bernhard Schölkopf, 'Elements of causal inference: Foundations and learning algorithms', (2017).

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

[25] Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, et al., 'On causally disentangled representations', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8089–8097, (2022).

[26] Ali Lotfi Rezaabad and Sriram Vishwanath, 'Learning representations by maximizing mutual information in variational autoencoders', in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2729–2734. IEEE, (2020).

[27] Danilo Rezende and Shakir Mohamed, 'Variational inference with normalizing flows', in *International conference on machine learning*, pp. 1530–1538. PMLR, (2015).

[28] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang, 'Weakly supervised disentangled generative causal representation learning', *Journal of Machine Learning Research*, **23**, 1–55, (2022).

[29] Robert Sicks, Ralf Korn, and Stefanie Schwaar, 'A generalised linear model framework for variational autoencoders based on exponential dispersion families', *J. Mach. Learn. Res.*, **22**, 233:1–233:41, (2020).

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

[31] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer, 'Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness', in *International Conference on Machine Learning*, pp. 6056–6065. PMLR, (2019).

[32] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro, 'Challenges in disentangling independent factors of variation', *arXiv preprint arXiv:1711.02245*, (2017).

[33] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer, 'On disentangled representations learned from correlated data', in *International Conference on Machine Learning*, pp. 10401–10412. PMLR, (2021).

[34] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis, 'The max-min hill-climbing bayesian network structure learning algorithm', *Machine learning*, **65**(1), 31–78, (2006).

[35] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar, 'Decaf: Generating fair synthetic data using causally-aware generative networks', *Advances in Neural Information Processing Systems*, **34**, 22221–22233, (2021).

[36] Bingyang Wen, Luis Oliveros Colon, KP Subbalakshmi, and Rajarathnam Chandramouli, 'Causal-tgan: Generating tabular data using causal generative adversarial networks', *arXiv preprint arXiv:2104.10680*, (2021).

[37] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu, 'Achieving causal fairness through generative adversarial networks', in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, (2019).

[38] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni, 'Modeling tabular data using conditional gan', *Advances in Neural Information Processing Systems*, **32**, (2019).

[39] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang, 'Causalvae: Disentangled representation learning via neural structural causal models', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, (2021).

[40] Yue Yu, Jie Chen, Tian Gao, and Mo Yu, 'Dag-gnn: Dag structure learning with graph neural networks', in *International Conference on Machine Learning*, pp. 7154–7163. PMLR, (2019).

[41] Zilong Zhao, Aditya Kunar, Hiek van der Scheer, Robert Birke, and Lydia Yiyu Chen, 'Ctab-gan: Effective table data synthesizing', *ArXiv*, **abs/2102.08369**, (2021).

[42] Zhilin Zheng and Li Sun, 'Disentangling latent space for vae by label relevant/irrelevant dimensions', *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12184–12193, (2018).

[43] Seunghwan An, Kyungwoo Song, and Jong-June Jeon, 'Causally disentangled generative variational autoencoder', *arXiv preprint arXiv:2302.11737*, (2023).