Journal of Machine Learning Research 23 (2022) 1-55 Submitted 1/21; Revised 7/22; Published 7/22

Weakly Supervised Disentangled Generative Causal Representation Learning

Xinwei Shen XSHENAL@UST.HK
Department of Mathematics
The Hong Kong University of Science and Technology
Hong Kong, China

Furui Liu LIUFURUI@ZHEJIANGLAB.COM
Zhejiang Laboratory
Hangzhou, China

Hanze Dong HDONGAJ@UST.HK
Department of Mathematics
The Hong Kong University of Science and Technology
Hong Kong, China

Qing Lian QLIANAB@UST.HK
Department of Computer Science
The Hong Kong University of Science and Technology
Hong Kong, China

Zhitang Chen CHENZHITANG2@HUAWEI.COM
Huawei Noah's Ark Lab
Shenzhen, China

Tong Zhang TONGZHANG@UST.HK
Department of Computer Science and Mathematics
The Hong Kong University of Science and Technology
Hong Kong, China

Editor: Yoshua Bengio

Abstract

This paper proposes a Disentangled gEnerative cAusal Representation (DEAR) learning method under appropriate supervised information. Unlike existing disentanglement methods that enforce independence of the latent variables, we consider the general case where the underlying factors of interests can be causally related. We show that previous methods with independent priors fail to disentangle causally related factors even under supervision. Motivated by this finding, we propose a new disentangled learning method called DEAR that enables causal controllable generation and causal representation learning. The key ingredient of this new formulation is to use a structural causal model (SCM) as the prior distribution for a bidirectional generative model. The prior is then trained jointly with a generator and an encoder using a suitable GAN algorithm incorporated with supervised information on the ground-truth factors and their underlying causal structure. We provide theoretical justification on the identifiability and asymptotic convergence of the proposed method. We conduct extensive experiments on both synthesized and real data sets to demonstrate the effectiveness of DEAR in causal controllable generation, and the bene-

©2022 Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-0080.html.

fits of the learned representations for downstream tasks in terms of sample efficiency and distributional robustness.

Keywords: disentanglement, causality, representation learning, deep generative model

1. Introduction

Consider the observed data x from a distribution q_x on $\mathcal{X} \subseteq \mathbb{R}^d$ and the latent variable z from a prior p_z on $\mathcal{Z} \subseteq \mathbb{R}^k$. In bidirectional generative models (BGMs), we are normally interested in learning an *encoder* $E : \mathcal{X} \to \mathcal{Z}$ to infer latent variables and a *generator* $G : \mathcal{Z} \to \mathcal{X}$ to generate data, to achieve both representation learning and data generation. Classical BGMs include Variational Autoencoder (VAE) (Kingma and Welling, 2014) and BiGAN (Donahue et al., 2017; Dumoulin et al., 2017). In representation learning, it was argued that an effective representation for downstream learning tasks should disentangle the underlying factors of variation (Bengio et al., 2013). In generative modeling, it is highly desirable if one can control the semantic generative factors by aligning them with the latent variables such as in StyleGAN (Karras et al., 2019). Both goals can be achieved with the disentanglement of latent variable z, which informally means that each dimension of z measures a distinct factor of variation in the data (Bengio et al., 2013).

Earlier unsupervised disentanglement methods mostly regularized the VAE objective to encourage independence of learned representations (Higgins et al., 2017; Burgess et al., 2017; Kim and Mnih, 2018; Chen et al., 2018; Kumar et al., 2018). Later, Locatello et al. (2019) showed that unsupervised learning of disentangled representations is impossible: many existing unsupervised methods are brittle, requiring careful supervised hyperparameter tuning or implicit inductive biases. To promote identifiability, recent work resorted to various forms of supervision (Locatello et al., 2020b; Shu et al., 2020; Locatello et al., 2020a). In this work, we also incorporate supervision on the ground-truth factors in the form of a certain number of annotated labels as described in Section 3.2. We will present experimental results showing that our method remains competitive with a small amount of labeled data (a minimum of around 100 samples).

Most of the existing methods, including those mentioned above, are built on the assumption that the underlying factors of variation are mutually independent. However, in many real-world cases, the semantically meaningful factors of interests are not independent (Bengio et al., 2020). Instead, such high-level variables are often causally related, i.e., connected by a causal graph.

In this paper, we prove formally that methods with independent priors fail to disentangle causally related factors. Motivated by this observation, we propose a new method to learn disentangled generative causal representations called DEAR. The key ingredient of our formulation is a structural causal model (SCM) (Pearl et al., 2000) as the prior for latent variables in a bidirectional generative model. As discussed in Section 4.1.2, we assume that a super-graph of the underlying causal graph is known a priori, which ranges from the causal ordering of the nodes in the graph to the true causal structure. The causal model prior is then learned jointly with a generator and an encoder using a suitable GAN (Goodfellow et al., 2014) algorithm. Moreover, we establish theoretical guarantees for DEAR on how it resolves the unidentifiability issue of many existing methods as well as on the asymptotic convergence of the proposed algorithm.

An immediate application of DEAR is causal controllable generation, which can generate data from many desired interventional distributions of the latent factors. Another useful application of disentangled representations is to use such representations in downstream tasks, leading to better sample complexity (Bengio et al., 2013; Schölkopf et al., 2012). Moreover, it is believed that causal disentanglement is invariant and thus robust under distribution shifts (Schölkopf, 2019; Arjovsky et al., 2019). In this paper, we demonstrate these conjectures in various downstream prediction tasks for the proposed DEAR method, which has theoretically guaranteed disentanglement property.

We summarize our main contributions as follows:

- We formally identify a problem with previous disentangled representation learning methods using the independent prior assumption, and prove that they fail to disentangle when the underlying factors of interests are causally related, even under supervision of the latents.
- We propose a new disentangled learning method, DEAR, which integrates an SCM prior into a bidirectional generative model, trained with a suitable GAN algorithm.
- We provide theoretical justification on the identifiability[1] of the proposed formulation and the asymptotic convergence of our algorithm.
- Extensive experiments are conducted on both synthesized and real data to demonstrate the effectiveness of DEAR in causal controllable generation, and the benefits of the learned representations for downstream tasks in terms of sample efficiency and distributional robustness.

Notation Throughout the paper, all distributions are assumed to be absolutely continuous with respect to Lebesgue measure unless indicated otherwise. For a vector x, let $[x]_i$ denote the i-th component of x. For a scalar function $h(x, y)$, let $\nabla_x h(x, y)$ denote its gradient with respect to x and $\nabla_x^2 h(x, y)$ denote its Hessian matrix with respect to x. For a vector function $g(x, y)$, let $\nabla_x g(x, y)$ denote its Jacobian matrix with respect to x. Without ambiguity, ∇_x is denoted by ∇ for simplicity. Notation $\|\cdot\|$ stands for the Euclidean norm.

Definition 1 (Smoothness) *Consider a function $h(x) : \mathbb{R}^d \to \mathbb{R}$. $h(x)$ is ℓ_0-smooth with respect to x if $h(x)$ is differentiable and its gradient is ℓ_0-Lipschitz continuous, i.e., we have*

$$\|\nabla h(x) - \nabla h(x')\| \leq \ell_0 \|x - x'\|, \quad \forall x, x' \in \mathbb{R}^d.$$

Definition 2 (Polyak-Łojasiewicz) *For a set $\mathcal{S} \subseteq \mathbb{R}^d$, consider a function $h(x) : \mathcal{S} \to \mathbb{R}$ and let $h^* = \min_{x \in \mathcal{S}} h(x)$. Then $h(x)$ satisfies the Polyak-Łojasiewicz (PL) condition if there exists $c > 0$ such that for all $x \in \mathcal{S}$*

$$h(x) - h^* \leq c\|\nabla h(x)\|_2^2.$$

1. Note that the identifiability in this work differs from that in Khemakhem et al. (2020) in terms of goals and assumptions. See more discussions in the related work and below Proposition 5.

Roadmap In Section 2, we discuss the related work. In Section 3, we introduce the problem setting of disentangled generative causal representation learning and identify a problem with previous methods. In Section 4, we propose the model, formulation and algorithm of DEAR, and provide theoretical justifications on both identifiability and asymptotic convergence. We then present empirical studies concerning causal controllable generation, downstream tasks and structure learning as well as ablation studies in Section 5, and conclude in Section 6. Detailed proofs of all theorems, propositions and lemmas are deferred to Appendix A.

2. Related work

VAE-based disentanglement methods. A number of methods have been proposed to enrich the VAE loss by various regularizers to enforce the independence of the latent variables. β-VAE (Higgins et al., 2017) and Annealed VAE (Burgess et al., 2017) introduced extra constraints on the capacity of the latent bottleneck by adjusting the role of the KL term; Factor-VAE (Kim and Mnih, 2018) and β-TCVAE (Chen et al., 2018) encouraged the aggregated posterior (i.e., the marginal distribution of $E(x)$) to be factorized by penalizing its total correlation; DIP-VAE (Kumar et al., 2018) enforced a factorized aggregated posterior differently by matching its moments with those of a factorized prior. Going beyond the independence perspective, Suter et al. (2019) considered disentangled causal mechanisms, meaning that all the generative factors are conditionally independent given a common confounder. This is one special case of causal relationship, while we consider more general cases where the factors can have more complex causal relationships, e.g., one factor can be a direct cause of another one.

Based on the above methods, Locatello et al. (2020b) and Locatello et al. (2020a) further incorporated supervised information on a few labels of the generative factors and pairs of observations which differ by a few factors respectively, where the former is more related to ours which is discussed detailedly in Section 3.2. Shu et al. (2020) proposed several concepts related to disentanglement, based on which they analyzed three forms of weak supervision including restricted labeling, match pairing, and rank pairing.

Going beyond the independent prior, Khemakhem et al. (2020) proposed a conditional VAE where the latent variables are assumed to be conditionally independent given some additionally observed variables. Built upon developments of nonlinear ICA, they presented the first principled identifiability theory of latent variable models, in particular VAEs, thus leading to a form of provable disentanglement under suitable conditions. Our work, in contrast, does not aim at achieving general identifiability of latent variable models or general provable disentanglement, but contributes to resolving the failure of existing methods in disentangling causally related factors. With this motivation, we consider more general model assumptions on the latent structure as well as generating transformations than those in Khemakhem et al. (2020) which apply more suitably to real-world data. To achieve disentanglement of causal factors, we need to adopt a more direct and somehow stronger form of supervision than Khemakhem et al. (2020), i.e., we require annotated labels of true factors for a possibly small number of samples. See Appendix C for a discussion on the two forms of supervision. The model in Khemakhem et al. (2020), however, has not yet been applied with the most advanced network architecture for image generation such as

StyleGAN (Karras et al., 2019), nor can their conditional independent prior models the causal structure of true factors. Therefore, their model and theory do not apply here and our work should be regarded complementary.

To avoid the unidentifiability of the standard Gaussian prior caused by rotation transformations, Stühmer et al. (2020) proposed hierarchical non-Gaussian priors for unsupervised disentanglement, which is not rotationally invariant. However, there remains other kinds of mixing transformations that leave these priors invariant, leading to unidentifiability. Besides, their proposed priors cannot model the causal relationships.

Recently, a concurrent work by Träuble et al. (2021) conducted a large-scale empirical study to investigate the behavior of the most prominent disentanglement approaches on correlated data. In particular, they considered the case where the ground-truth factors exhibit pairwise correlation. Although pairwise correlation largely generalizes the independence assumption, it is less general than the causal correlation that we consider. For example, a parental node with multiple children immediately goes beyond pairwise correlation. Moreover, Träuble et al. (2021) focused on verifying the problem that existing methods fail to learn disentangled representations for strongly correlated factors, while we identify the problem as a motivation to propose a method to resolve it and learn disentangled representations under the causal case.

GAN-based disentanglement methods. Existing GAN-based methods, including InfoGAN (Chen et al., 2016) and InfoGAN-CR (Lin et al., 2020), differed from our proposed formulation mainly in two folds. First, they still assumed an independent prior for latent variables, so suffered from the same problem with the previous VAE-based methods mentioned above. Besides, the idea of InfoGAN-CR was to encourage each latent code to make changes that are easy to detect, which only applies well when the underlying factors are independent. Second, as a bidirectional generative modeling method, InfoGAN further required variational approximation apart from adversarial training, which is inferior to the principled formulation in BiGAN and AGES (Shen et al., 2020) that we adopt.

Generative modeling involving causal models in the latent space. CausalGAN (Kocaoglu et al., 2018) and a concurrent work (Moraffah et al., 2020) of ours, were unidirectional generative models (i.e., a generative model that learns a single mapping from the latent variable to data) that build upon a cGAN (Mirza and Osindero, 2014). They assigned an SCM to the conditional attributes while leaving the latent variables as independent Gaussian noises. The limit of a cGAN is that it always requires full supervision on attributes to apply conditional adversarial training. Also, the ground-truth factors were directly fed into the generator as the conditional attributes, without any extra effort to align the dimensions between the latent variables and the underlying factors, so their models had nothing to do with disentanglement learning. Moreover, their unidirectional nature made them unable to learn representations. Besides, they only considered binary factors, so the consequent semantic interpolations appear non-smooth, as shown in Appendix G.

CausalVAE (Yang et al., 2021) assigned the SCM directly on the latent variables, while built upon iVAE (Khemakhem et al., 2020), it adopted a conditional prior given the ground-truth factors so was also limited to a fully supervised setting.

GraphVAE (He et al., 2018) generalized the chain-structured latent space proposed in Ladder VAE (Sønderby et al., 2016) and imposed an SCM into the latent space of VAE.

The motivation behind GraphVAE is to improve the expressive capacity of VAE rather than to disentangle the underlying causal factors as ours. Purely from observational data and without any supervision on the underlying factors, the impossibility result from Locatello et al. (2019) indicated that a VAE model cannot identify the true factors. Therefore, the representations learned by GraphVAE were not guaranteed to disentangle the generative factors, and consequently the learned SCM did not reflect the true causal structure in principle. Moreover, their adopted VAE loss (ELBO) required an explicit form of KL divergence between the prior and the posterior, which limited the model choice for the SCM. Specifically, GraphVAE used an additive noise model with Gaussian noises. In contrast, our method does not require the distribution induced by the SCM to be explicitly expressed and in principle allows any SCMs that can be reparametrized as a generative model (i.e., given the exogenous noises, one can generate all the variables by ancestral sampling). For comparison, in our experiments, we include a baseline which extends the original GraphVAE method to incorporate the same amount of supervision as ours.

Generative modeling involving other structured latent spaces. VLAE (Zhao et al., 2017) decomposed the latent space into separate chunks each of which is processed at different levels of the encoder and decoder. VQ-VAE-2 (Razavi et al., 2019) used a two-level latent space along with a multi-stage generation mechanism to capture both high and low level information of data. SAE (Leeb et al., 2020) encouraged a hierarchical structure in the latent space through the structural architecture of the decoder. These methods essentially adopted implicit probabilistic or architectural hierarchies, in contrast to the causal structure that we impose to the latent space, and thus cannot achieve the goal of causal disentanglement. For example, the hierarchy in SAE represents the level of abstraction, in the sense that more high-level, abstract features are processed deeper in the decoder and low-level, linear features are treated towards the end of the network. Such hierarchy differs essentially from the causal structure that we consider.

Other works considered inferring the latent causal structure from visual data in the reinforcement learning setting (Dasgupta et al., 2019; Nair et al., 2019). In particular, Nair et al. (2019) developed learning-based approaches to induce causal knowledge in the form of directed acyclic graphs, which was then utilized in learning goal-conditioned policies. The interactive environment enables the agent to perform actions and observe their outcomes. Therefore, the resulting data involves various interventions each of which entails an SCM and thus is essentially different from the common setting in the disentanglement literature which is also considered in this paper, where the observed data are independent and identically distributed.

3. Problem setting

In this section, we describe the probabilistic framework of disentanglement learning based on bidirectional generative models (BGMs) with supervision, and formalize the unidentifiability problem with previous methods.

3.1 Generative model

We follow the commonly assumed two-step data generating process that first samples the underlying generative factors, and then conditional on those factors, generates the data (Kingma and Welling, 2014). During the generation process, the generator induces the generated conditional $p_G(x|z)$ and generated joint distribution $p_G(x,z) = p_z(z)p_G(x|z)$. During the inference process, the encoder induces the encoded conditional $q_E(z|x)$ which can be a factorized Gaussian and the encoded joint distribution $q_E(x,z) = q_x(x)q_E(z|x)$.

We consider the following objective for generative modeling:

$$L_{\text{gen}}(E,G) = D_{\text{KL}}(q_E(x,z), p_G(x,z)), \tag{1}$$

where $D_{\text{KL}}(q,p) = \int q(x,z)\log(q(x,z)/p(x,z))dxdz$ is the Kullback-Leibler (KL) divergence between two distributions. Objective (1) is shown to be equivalent to the negative evidence lower bound (ELBO),

$$\mathbb{E}_{x\sim q_x}[-\mathbb{E}_{q_E(z|x)}\log p_G(x|z) + D_{\text{KL}}(q_E(z|x), p_z(z))], \tag{2}$$

used in VAEs up to a constant, and ELBO allows a closed form to be optimized easily only with factorized Gaussian prior, encoder and generator (Shen et al., 2020).

Since constraints on the latent space are required to enforce disentanglement, it is desirable that the distribution families of $q_E(x,z)$ and $p_G(x,z)$ should be large enough, especially for complex data like images. As demonstrated in literature on image generation (Karras et al., 2019; Mescheder et al., 2017), implicit distributions, where the randomness is fed into the input or intermediate layers of the network, are favored over factorized Gaussians in terms of expressiveness. Then minimizing (1) requires adversarial training, as discussed detailedly in Section 4.3.

3.2 Supervised regularizer

To guarantee disentanglement, we incorporate supervision when training the BGM. The first part of supervision consists of a certain number of annotated labels of the ground-truth factors, following the similar idea in Locatello et al. (2020b) but with a different formulation. We leverage another part of supervision on the graph structure of the factors, which will be discussed in Section 4.1.2. Specifically, let $\xi \in \mathbb{R}^m$ be the underlying ground-truth factors of interests of data x, following distribution p_ξ, and $[y]_i$ be some continuous or discrete annotated observation of the i-th underlying factor $[\xi]_i$, satisfying $[\xi]_i = \mathbb{E}([y]_i|x)$ for $i = 1,\dots,m$. For example, in the case of human face images, $[y]_1$ can be the binary label indicating whether a person is young or not, and $[\xi]_1 = \mathbb{E}([y]_1|x) = \mathbb{P}([y]_1 = 1|x)$ is the probability of being young given one image x.

Let $\bar{E}(x)$ be the deterministic part of the stochastic transformation $E(x)$, i.e., $\bar{E}(x) = \mathbb{E}(E(x)|x)$ by integrating out the additional randomness injected into the encoder, which is used for representation learning. For instance, consider a Gaussian encoder satisfying $E(x)|x \sim \mathcal{N}(m(x), \Sigma(x))$ which can be reparametrized by $E(x) = m(x) + \Sigma(x)^\top\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. Then the deterministic part is the mean, i.e., $\bar{E}(x) = m(x)$.

We consider the following objective:

$$L(E,G) = L_{\text{gen}}(E,G) + \lambda L_{\text{sup}}(E), \tag{3}$$

where the supervised regularizer is $L_{\text{sup}} = \mathbb{E}_{x,y}[l_s(E; x, y)]$ with $l_s = \sum_{i=1}^{m} \text{CE}([\bar{E}(x)]_i, [y]_i)$ if $[y]_i$ is the binary or bounded (and normalized to $[0, 1]$) continuous label of factor $[\xi]_i$, where $\text{CE}(l, y) = -y \log \sigma(l) - (1 - y) \log(1 - \sigma(l))$ is the cross-entropy loss with $\sigma(\cdot)$ being the sigmoid function; $l_s = \sum_{i=1}^{m} ([\bar{E}(x)]_i - [y]_i)^2$ if $[y]_i$ is the continuous observation of $[\xi]_i$. $\lambda > 0$ is the coefficient to balance both terms. Through ablation studies in Section 5.4, we empirically find the choice of λ insensitive to different tasks and data sets, and hence set $\lambda = 5$ in all experiments.

Note that in the objective (3), the unsupervised generative modeling loss and the supervised regularizer are decoupled in terms of taking expectations, in contrast to the conditional GANs where supervised labels are involved in the GAN loss. This enables one to use two separate samples with different sample sizes to estimate the two terms in (3) during training. Since in practice we may only have access to a limited amount of annotated labels, this property makes the formulation applicable in such semi-supervised settings. In the experiments, we conduct ablation studies to investigate how our method performs with varying amounts of labeled samples available.

In addition, Locatello et al. (2020b) propose a regularizer $L_{\text{sup}} = \sum_{i=1}^{m} \mathbb{E}_{x,z}(\text{CE}([\bar{E}(x)]_i, [z]_i))$ involving only the latent variable z which is a part of the generative model, without distinguishing the model component z from the ground-truth factor ξ and its observation y. Hence they do not establish formal theoretical justification on disentanglement. Moreover, they follow the earlier VAE-based methods to adopt a VAE loss (2) for generative modeling with an independent prior and an additional regularizer to enforce independence of the latent variables, which suffers from the unidentifiability problem described in the next section.

3.3 Unidentifiability with an independent prior

Intuitively, the above supervised regularizer aims at ensuring some kind of alignment between the underlying factor ξ and the latent variable z in the model. We start with the definition of a disentangled representation following this intuition.

Definition 3 (Disentangled representation) *Given the underlying factor $\xi \in \mathbb{R}^m$ of data x, a deterministic encoder E is said to learn a disentangled representation with respect to ξ if $\forall i = 1, \ldots, m$, there exists a 1-1 function g_i such that $[E(x)]_i = g_i([\xi]_i)$. Further, a stochastic encoder E is said to be disentangled with respect to ξ if its deterministic part $\bar{E}(x)$ is disentangled with respect to ξ.*

Note that in general, the goal of disentanglement allows for permutations in the ground-truth factors. For example one may expect for all i there exists j which is not necessarily equal to i such that $[E(x)]_i = g_j([\xi]_j)$. However since in our method we supervise each latent dimension by the annotated label of each ground-truth factor, we can expect a component-wise correspondence between $E(x)$ and ξ, as justified formally in Proposition 5 below.

As introduced above, we consider the general case where the underlying factors of interests are causally related. Then the goal becomes to disentangle the causal factors. Previous methods mostly use an independent prior for z, which contradicts the truth. We make this formal through the following proposition, which indicates that the disentangled representation is generally unidentifiable with an independent prior.

Proposition 4 *Let E^* be any encoder that is disentangled with respect to ξ. Let $b^* = L_{\text{sup}}(E^*)$, $a = \min_G L_{\text{gen}}(E^*, G)$, and $b = \min_{\{(E,G):L_{\text{gen}}=0\}} L_{\text{sup}}(E)$. Assume the elements of ξ are connected by a causal graph whose adjacency matrix A_0 is not a zero matrix. Suppose the prior p_z is factorized, i.e., $p_z(z) = \prod_{i=1}^k p_i([z]_i)$. Then we have $a > 0$, and either when $b^* \geq b$ or $b^* < b$ and $\lambda < \frac{a}{b-b^*}$, there exists a solution (E', G') so that E' is entangled and for any generator G, we have $L(E', G') < L(E^*, G)$.*

This proposition directly suggests that minimizing (3) favors an entangled solution (E', G') over the one with a disentangled encoder E^*. Thus, with an independent prior we have no way to identify the disentangled solution with λ that is not large enough. However, in real applications, it is impossible to estimate the threshold, and too large λ makes it difficult to learn the BGM. After our work was submitted, we were brought attention to a theoretical result in Träuble et al. (2021) that is similar to our Proposition 4. A discussion on the two independently proposed results is given in Appendix A.2 after the proof. In the following section, we propose a solution to this problem.

4. Causal disentanglement learning

In this section, we propose the DEAR method for causal disentanglement learning. We start with an introduction to the model structure in Section 4.1. Then we present the formulation of DEAR as well as its identifiability of disentanglement at a population level in Section 4.2. The DEAR algorithm is described in Section 4.3 with its consistency results established in Section 4.4.

4.1 Generative model with a causal prior

We introduce the proposed bidirectional generative model with a causal model prior, and discuss the learning of the adjacency matrix. Based on the model we describe the mechanism of causal controllable generation from interventional distributions. We further propose a composite prior to deal with the issue of setting the latent dimension.

4.1.1 SCM prior

We propose to use a causal model as the prior p_z. Specifically we adopt the general nonlinear Structural Causal Model (SCM) proposed by Yu et al. (2019) as follows

$$z = f((I - A^\top)^{-1} h(\epsilon)) := F_\beta(\epsilon), \tag{4}$$

where A is the weighted adjacency matrix of the directed acyclic graph (DAG) upon the k elements of z (i.e., $A_{ij} \neq 0$ if and only if $[z]_i$ is the parent of $[z]_j$), ϵ denotes the exogenous variables following $\mathcal{N}(0, I)$, f and h are element-wise transformations that are generally nonlinear, and $\beta = (f, h, A)$ denotes the set of parameters of f, h and A, with the parameter space $\mathcal{B}$. Further let $\mathbf{I}_A = \mathbf{I}(A \neq 0)$ denote the corresponding binary adjacency matrix, where $\mathbf{I}(\cdot)$ is the element-wise indicator function.

When f is invertible, (4) is equivalent to

$$f^{-1}(z) = A^\top f^{-1}(z) + h(\epsilon), \tag{5}$$

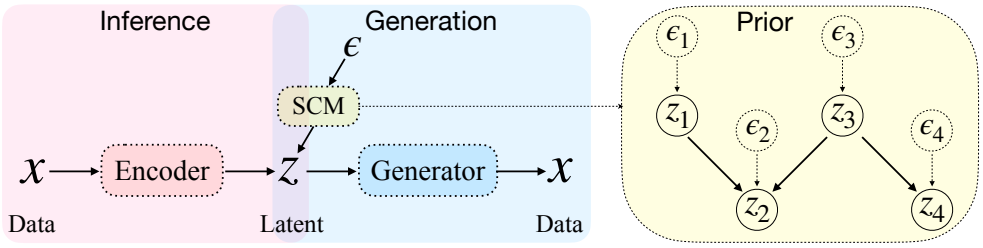


Figure 1: Model structure of a BGM (left) with an SCM prior (right).

which indicates that the factors z satisfy a linear SCM after nonlinear transformation f, and enables interventions on latent variables as discussed later.

By combining the above SCM prior and the encoder and generator introduced in Section 3.1, we end up with the model structure presented in Figure 1. Note that different from our model where z is the latent variable following the prior (4) with the goal of causal disentanglement, Yu et al. (2019) propose a causal discovery method where variables z in SCM (4) are observed with the aim of learning the causal structure among z.

4.1.2 Learning of A

In causal structure learning, the graph is required to be acyclic. Traditional causal discovery methods such as PC (Spirtes et al., 2000) or GES (Chickering, 2002) deal with the combinatorial problem over the discrete space of DAGs. Recently, Zheng et al. (2018) proposed an equality constraint whose satisfaction ensures acyclicity and solved the problem with the augmented Lagrangian method, which however leads to optimization difficulties (Ng et al., 2020). In addition, identifiability of the causal structure from purely observational data is known as an important issue in causal discovery. Despite a number of results on structure identifiability under various parametric or semi-parametric assumptions (Zhang and Hyvarinen, 2009; Peters and Bühlmann, 2014), in a general nonparametric setting, however, it cannot be guaranteed. Yu et al. (2019) did not discuss the identifiability of the SCM (4) under general cases.

In many problems of disentanglement, we have some prior information on the causal structure of the factors of interests based on common knowledge or expertise. In particular, we may know a causal ordering of the factors. In addition to the ordering, for some factors, we may know that one particular factor cannot be a direct cause of another one, which helps us remove some redundant edges in advance. Therefore, in this paper with the focus on disentanglement, we utilize such prior information on the graph structure in disentanglement learning and leave incorporating causal discovery from scratch to future work. Formally, we assume the super-graph of the true binary graph $\mathbf{I}_{A_0}$ is given, the best case of which is the true graph while the worst is that only the causal ordering is available. Then we learn the weights of the non-zero elements of the prior adjacency matrix that indicate the sign and scale of causal effects, jointly with other parameters of the generative model using the formulation and algorithm described in Sections 4.2 and 4.3.

As discussed in Section 4.2, such prior knowledge makes the structure identifiability easy to hold. Moreover, the given super-graph ensures the acyclicity of the adjacency matrix,

allowing us to get rid of the additional acyclicity constraint. In Section 5.3, we investigate how our method performs in learning the graph structure and weighted adjacency given various amounts of prior graph information. Note that even when a super-graph is available, to our best knowledge, no previous disentanglement method except GraphVAE (He et al., 2018) can utilize them to disentangle causal factors with guarantee, but we propose one such method and show its effectiveness. In fact, He et al. (2018) also assumed an ordering over the latent nodes by specifying that the parents of node $z_i, i = 1, \ldots, k-1$ come from the set $\{z_{i+1}, \ldots, z_k\}$. Later experiments suggest that GraphVAE shows inferior performance compared with ours.

4.1.3 Generation from interventional distributions

One immediate application of our proposed model is causal controllable generation from interventional distributions of the latent variables. We now describe the mechanism. To enable intervention under SCM (5), we require f to be invertible. Then interventions can be formalized as operations that modify a subset of equations in (5) (Pearl et al., 2000).

Suppose we would like to intervene on the i-th dimension of z, i.e., $\mathrm{Do}([z]_i = c)$, where c is a constant. Once we obtain the latent factors z inferred from data x, i.e., $z = E(x)$, or sampled from prior p_z, we follow the modified equations in (5) to obtain z' on the left-hand side using ancestral sampling by performing (5) iteratively, where ϵ can be either fixed or resampled from its prior. Then we decode the latent factor z' that follows the given interventional distribution to generate the desired sample $G(z')$. In Section 5.1 we define the two types of interventions of most interests in applications. We discuss how our method generalizes to unseen interventions in Appendix D.

4.1.4 Latent dimension and composite prior

Another issue of the model is how to set the latent dimension k of the generative model, to handle which we propose the so-called composite prior. Recall that m is the number of generative factors that we are interested to disentangle, for example, all the semantic concepts related to some filed, where m tends to be smaller than the total number M of generative factors. The latent dimension k should be no less than M to allow a sufficient degree of freedom in order to generate or reconstruct data well. Since M is generally unknown in reality, we set a sufficiently large k, at least larger than m which is a trivial lower bound of M.

Then we propose to use a prior that is a composition of a causal model for the first m dimensions and another distribution for the other $k-m$ dimensions to capture other factors necessary for generation, like a standard Gaussian. In this way the first m dimensions of z aim at learning the disentangled representation of the m factors of interests, while the role of the remaining $k-m$ dimensions is to capture other factors that are necessary for generation whose structure is neither cared nor explicitly modeled. Under this model framework, we do not require the availability of annotated labels for all generative factors of data, but only the ones of our interests to disentangle are used in the supervised regularizer in (3), which broadens the applications of our method.

4.2 DEAR formulation

In this section, we first present the formulation of DEAR. Compared with the BGM described in Section 3.1, now we have one more module to learn which is the SCM prior. Thus $p_G(x,z)$ becomes $p_{G,F}(x,z) = p_F(z)p_G(x|z)$ where $p_F(z)$ is the distribution of $F_\beta(\epsilon)$ with $\epsilon \sim \mathcal{N}(0,I)$. We then rewrite the generative model loss as follows

$$L_{\text{gen}}(E,G,F) = D_{\text{KL}}(q_E(x,z), p_{G,F}(x,z)). \tag{6}$$

Then we propose the following formulation to learn disentangled generative causal representations:

$$\min_{E,G,F} \; L(E,G,F) := L_{\text{gen}}(E,G,F) + \lambda L_{\text{sup}}(E). \tag{7}$$

Now we show the identifiability of disentanglement of DEAR in contrast to the unidentifiability result in Proposition 4. Proposition 5 indicates that under appropriate conditions, the DEAR formulation (7) at a population level can learn the disentangled representations defined in Definition 3. Here, Assumption 1 supposes a sufficiently large capacity of the SCM in (4) to contain the underlying distribution p_ξ, which is reasonable due to the generalization of the nonlinear SCM.

Assumption 1 *The underlying distribution p_ξ belongs to the distribution family $\{p_\beta : \beta \in \mathcal{B}\}$, i.e., there exits $\beta_0 = (f_0, h_0, A_0)$ such that $p_\xi = p_{\beta_0}$.*

Proposition 5 (Identifiability) *Assume the infinite capacity of E and G and Assumption 1. Let $(E^*, G^*, F^*) \in \operatorname{argmin}_{E,G,F} L(E,G,F)$ which is the solution of DEAR formulation (7). Then E^* is disentangled with respect to ξ as defined in Definition 3.*

Note that Proposition 5 states the identifiability at the population level, i.e., the loss function is taken the expectation over distributions of both the data and labels of the true factors. Thus we clarify that Proposition 5 does not obtain general provable disentanglement which should be analyzed with a much weaker form of supervision on the true factors, e.g., as in Khemakhem et al. (2020). In contrast, the specific identifiability stated in Proposition 5 should be interpreted as a counterpart of the unidentifaibility result in Proposition 4. Specifically, Proposition 4 shows that the independent prior used by most existing disentanglement methods causes the contradiction between the generative loss L_{gen} and the supervised loss L_{sup} in (3), which makes the whole loss L prefer an entangled model. Therefore, even with the same amount of supervised labels of true factors, those methods cannot learn a generative model with disentangled latent representations. In contrast, Proposition 5 formally suggests that due to the introduction of the SCM prior, the two loss terms L_{gen} and L_{sup} in (7) can be simultaneously minimized and the jointly optimal solution leads to the disentangled model.

4.3 Algorithm

In this section, we propose the algorithm to solve the above formulation (7). Estimating L_{gen} requires the unlabeled data set $\{x_1, \ldots, x_N\}$ with sample size N, while estimating L_{sup} requires a labeled data set $\{(x_j, y_j) : j = 1, \ldots, N_s\}$, where the sample size N_s can be much

smaller than N. Without loss of generality, let $S_{\mathcal{G}} = \{x_1, \ldots, x_N, y_1, \ldots, y_{N_s}\}$ denote the training data set for the generative model.

We parametrize $E_\phi(x)$ and $G_\theta(z)$ by neural networks. As mentioned in Section 3.1, to enhance the expressiveness of the generative model, we use an implicit generated conditional $p_G(x|z)$, where we inject Gaussian noises to each convolution layer in the same way as Shen et al. (2020). Then the SCM prior $p_F(z)$ and implicit $p_G(x|z)$ make (6) lose an analytic form. Hence we adopt a GAN method to adversarially estimate the gradient of (6) as in Shen et al. (2020). Different from their setting, the prior also involves learnable parameters, that is, the parameters β of the SCM. In the following lemma we present the gradient formulas of (6).

Lemma 6 *Let $D^*(x,z) = \log[q_E(x,z)/p_{G,F}(x,z)]$. Then we have*

$$\begin{aligned}
\nabla_\theta L_{\text{gen}} &= -\mathbb{E}_{z\sim p_\beta(z)}[s(x,z)\nabla_x D^*(x,z)^\top|_{x=G_\theta(z)}\nabla_\theta G_\theta(z)], \\
\nabla_\phi L_{\text{gen}} &= \mathbb{E}_{x\sim q_x}[\nabla_z D^*(x,z)^\top|_{z=E_\phi(x)}\nabla_\phi E_\phi(x)], \\
\nabla_\beta L_{\text{gen}} &= -\mathbb{E}_\epsilon[s(x,z)(\nabla_x D^*(x,z)^\top \nabla_\beta G(F_\beta(\epsilon)) + \nabla_z D^*(x,z)^\top \nabla_\beta F_\beta(\epsilon))|_{z=F_\beta(\epsilon)}^{x=G(F_\beta(\epsilon))}],
\end{aligned} \tag{8}$$

where $s(x,z) = e^{D^*(x,z)}$ is the scaling factor.

Since D^* depends on the unknown densities, which makes the gradients in (8) uncomputable directly from data, we estimate the gradients by training a discriminator D via the empirical logistic regression:

$$\min_{D'} \frac{1}{N_d}\Bigg[\sum_{i:w_i=1}\log(1+e^{-D'(x_i,z_i)}) + \sum_{i:w_i=0}\log(1+e^{D'(x_i,z_i)})\Bigg], \tag{9}$$

where the class label $w_i = 1$ if $(x_i, z_i) \sim q_E$ and $w_i = 0$ if $(x_i, z_i) \sim p_{G,F}$, with $i = 1, \ldots, N_d$. We parametrize the discriminator using neural networks with parameter ψ.

Based on the above, we propose Algorithm 1 to learn disentangled generative causal representation.

4.4 Consistency

In this section, we show the asymptotic convergence of Algorithm 1. Let $\boldsymbol{\theta} = (\theta, \phi, \beta)$ denote the set of parameters of the generative model, where θ, ϕ and β denote the parameters of the generator, encoder and SCM prior respectively. According to such parametrization, we write the objective function in (7) as $L(\boldsymbol{\theta})$. In this section, we establish the consistency result of empirical estimator $\hat{\boldsymbol{\theta}}$, i.e., the output of Algorithm 1, under the parametric setting. Given a discriminator D, the approximate gradient used in the algorithm is denoted by

$$h_D(\boldsymbol{\theta}) = \begin{bmatrix} -\frac{1}{N}\sum_{i=1}^N \left[s(G_\theta(z_i), z_i)\nabla_x D(G_\theta(z_i), z_i)^\top \nabla_\theta G_\theta(z_i)\right] \\ \frac{1}{N}\sum_{i=1}^N \nabla_z D(x_i, E_\phi(x_i))^\top \nabla_\phi E_\phi(x_i) + \frac{\lambda}{N_s}\sum_{i=1}^{N_s}\nabla_\phi l_s(\phi; x_i, y_i) \\ -\frac{1}{N}\sum_{i=1}^N s(x,z)[\nabla_x D(x,z)^\top \nabla_\beta G(F_\beta(\epsilon_i)) + \nabla_z D(x,z)^\top \nabla_\beta F_\beta(\epsilon_i)]|_{z=F_\beta(\epsilon_i)}^{x=G(F_\beta(\epsilon_i))} \end{bmatrix}.$$

We first show in the following lemma that under appropriate conditions the approximate gradient $h_{\hat{D}}(\boldsymbol{\theta})$ based on the solution of (9) converges uniformly in probability to the true

Algorithm 1: Disentangled gEnerative cAusal Representation (DEAR) Learning

Input: training set $S_{\mathcal{G}}$, initial parameter $\phi, \theta, \beta, \psi$, batch-size n, meta-parameter T

for $t = 1, \ldots, T$ **do**
  for *multiple steps* **do**
    Sample $\{x_1, \ldots, x_n\}$ from the training set, $\{\epsilon_1, \ldots, \epsilon_n\}$ from $\mathcal{N}(0, I)$
    Generate from the causal prior $z_i = F_\beta(\epsilon_i), i = 1, \ldots n$
    Update ψ by descending the stochastic gradient:
    $\frac{1}{n}\sum_{i=1}^{n} \nabla_\psi \left[\log(1 + e^{-D_\psi(x_i, E_\phi(x_i))}) + \log(1 + e^{D_\psi(G_\theta(z_i), z_i)})\right]$
  Sample $\{x_1, \ldots, x_n, y_1, \ldots, y_{n_s}\}$, $\{\epsilon_1, \ldots, \epsilon_n\}$ as above; generate $z_i = F_\beta(\epsilon_i)$
  Compute θ-gradient: $-\frac{1}{n}\sum_{i=1}^{n} s(G_\theta(z_i), z_i)\nabla_\theta D_\psi(G_\theta(z_i), z_i)$
  Compute ϕ-gradient: $\frac{1}{n}\sum_{i=1}^{n} \nabla_\phi D_\psi(x_i, E_\phi(x_i)) + \frac{\lambda}{n_s}\sum_{i=1}^{n_s} \nabla_\phi l_s(\phi; x_i, y_i)$
  Compute β-gradient: $-\frac{1}{n}\sum_{i=1}^{n} s(G(z_i), z_i)\nabla_\beta D_\psi(G_\theta(F_\beta(\epsilon_i)), F_\beta(\epsilon_i))$
  Update parameters ϕ, θ, β using the gradients

Return: ϕ, θ, β

gradient. Recall the definition $D^*(x, z) = \log(q_E(x, z)/p_{G,F}(x, z))$ which depends on $\boldsymbol{\theta}$. Let $\mathcal{D}^* = \{D^*_{\boldsymbol{\theta}}(x, z) : \boldsymbol{\theta} \in \Theta\}$ denote the true discriminator class, and $\mathcal{D} = \{D(x, z)\}$ denote the modeled discriminator class with the norm $\|D\|_1 = \int |D(x, z)| p^*_{\boldsymbol{\theta}}(x, z) dx dz$, where $p^*_{\boldsymbol{\theta}}(x, z) = (q_E(x, z) + p_{G,F}(x, z))/2$ which induces the probability measure $\mu^*_{\boldsymbol{\theta}}$.

Lemma 7 *Assume the parameter space $\Theta = \{\boldsymbol{\theta} = (\theta, \phi, \beta)\}$ is compact. Assume the following regularity conditions hold:*

C1 $D^*_{\boldsymbol{\theta}}$ is smooth with respect to $\boldsymbol{\theta}$ over Θ, as defined in Definition 1.

C2 The modeled discriminator class $\mathcal{D}$ is compact, and contains the true class $\mathcal{D}^*$.

C3 $\{\mu^*_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is uniformly tight, i.e., for any $\epsilon > 0$, there exists a compact subset K_ϵ of $\mathcal{X} \times \mathcal{Z}$ such that for all $\boldsymbol{\theta} \in \Theta$, $\mu^*_{\boldsymbol{\theta}}(K_\epsilon) \geq 1 - \epsilon$.

C4 Functions in $\mathcal{D}$ have uniformly bounded function values, gradients and Hessians so that there exists a positive number $B_0 < \infty$ such that $\forall D \in \mathcal{D}$, $\forall x, z$, we have $|D(x, z)| \leq B_0$, $\|\nabla D(x, z)\| \leq B_0$ and $|tr(\nabla^2 D(x, z))| \leq B_0$.

C5 $\bar{E}_\phi$, ∇G_θ, ∇E_ϕ and ∇F_β are uniformly bounded.

C6 The training set for the discriminator is independent from that for the generative model.

Then there exists a sequence of $(N, N_s, N_d) \to \infty$ such that

$$\sup_{\boldsymbol{\theta} \in \Theta} \|h_{\hat{D}}(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta})\| \xrightarrow{p} 0, \tag{10}$$

where $\xrightarrow{p}$ means converging in probability.

Based on the above, we obtain the consistency of DEAR algorithm in the following theorem. It indicates that when the sample sizes grow large enough, with high probability, the DEAR algorithm approximately achieves the minimum of $L(\boldsymbol{\theta})$ which leads to the desired disentangled model according to Proposition 5.

Theorem 8 (Consistency) *Suppose the assumptions in Lemma 7 hold. Further assume the objective function $L(\boldsymbol{\theta})$ in (7) is smooth with respect to $\boldsymbol{\theta}$ and satisfies the Polyak-Łojasiewicz condition in Definition 2. Let $L^* = \min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$ Then there exists a sequence of $(N, N_s, N_d) \to \infty$ such that $L(\hat{\boldsymbol{\theta}}) \xrightarrow{p} L^*$.*

Remark. The Polyak-Łojasiewicz (PL) condition (Polyak, 1963) asserts that the suboptimality of a model is upper bounded by the norm of its gradient, which is a weaker condition than assumptions commonly made to ensure convergence, such as (strong) convexity. Recent literature showed that the PL condition holds for many machine learning scenarios including some deep neural networks (Charles and Papailiopoulos, 2018; Liu et al., 2020).

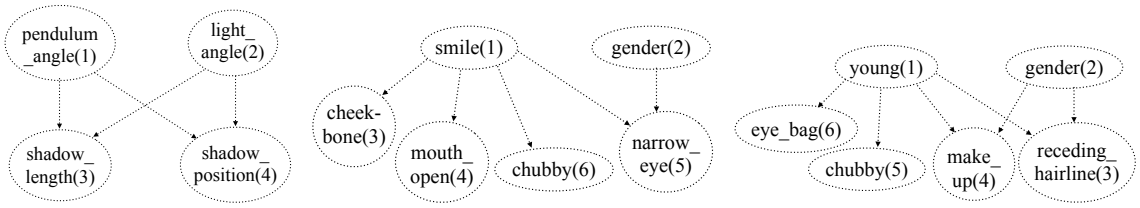
5. Experiments

We present the experimental studies in causal controllable generation in Section 5.1 which demonstrate the effectiveness of DEAR in causal disentanglement and support the theory in Section 4. Based on these theoretical and empirical justifications, we then apply the representations learned by DEAR in downstream prediction tasks in Section 5.2, and show the benefits of the disentangled causal representations in terms of sample efficiency and distributional robustness. In addition, we investigate the performance of DEAR in learning the causal structure and weighted adjacency of the SCM prior in Section 5.3. We also provide ablation studies in terms of varying regularization strength λ and various amounts of annotated labels in Section 5.4.[2]

We evaluate our methods on two data sets where the ground-truth generative factors are causally related, while most data sets used in previous disentanglement work are assumed or designed to have independent generative factors, for example, in the large scale experimental study by Locatello et al. (2019). The first data set that we use is a synthesized data set, Pendulum, similar to the one in Yang et al. (2021). As shown in Figure 3, each image is generated by four continuous factors: *pendulum_angle*, *light_angle*, *shadow_length* and *shadow_position* whose underlying structure is given in Figure 2(a) following physical mechanisms. To make the data set realistic, we introduce random noises when generating the two effects from the causes, representing the measurement error. We further introduce 20% corrupted data whose shadow is randomly generated, mimicking some environmental disturbance. The sample sizes for the training, validation and test set are all 6,724.

The second one is a real human face data set, CelebA (Liu et al., 2015), with 40 labeled binary attributes. Among them, we consider two groups of causally related factors of interests as shown in Figure 2(b,c). The sample sizes for the training, validation and test set are 162,770, 19,867, and 19,962. We believe these two data sets are diverse enough to assess our methods because they cover real and synthesized data, with continuous and discrete annotated labels. In addition, we test our method on benchmark data sets (Gondal et al., 2019) where the generative factors are independent. The results are given in Appendix E. All the details of the experimental setup, network architectures and the synthesized data set are given in Appendix F. Notably, all VAEs and DEAR use the same network architecture for the encoder and decoder (generator).

2. The code and data sets are available at `https://github.com/xwshen51/DEAR`.



(a) Pendulum (b) CelebA-Smile (c) CelebA-Attractive

Figure 2: Underlying causal structures.

5.1 Causal controllable generation

We first investigate the performance of our methods in disentanglement through applications in causal controllable generation. Traditional controllable generation methods mainly manipulate the independent generative factors (Karras et al., 2019), while we consider the general case where the factors are causally related. With a learned SCM as the prior, we are able to generate images from many desired interventional distributions of the latent factors. For example, we can manipulate only the cause factor while leaving its effects unchanged. Besides, the bidirectional framework presented in Figure 1 enables controllable generation either from scratch or a given unlabeled image.

We consider two types of interventions of most interests in applications. First, in traditional traversals, we manipulate one dimension of the latent vector while keeping the others fixed to either their inferred or sampled values (Higgins et al., 2017). A causal view of such operations is an intervention on all the variables by setting them as constants with only one of them varying. Another interesting type of interventional distribution is to intervene on only one latent variable, i.e., $\mathbb{P}_{\mathrm{do}([z]_i=c)}(z)$, and to observe how other variables change consequently. The proposed SCM prior enables us to conduct such interventions through the mechanism described in Section 4.1.3. One can naturally generalize it to intervene on more than one variable. For simplicity, we only present the results of intervening on one variable in the paper.

Figure 3-4 illustrate the results of causal controllable generation of the proposed DEAR method and the baseline method with independent priors, S-β-VAE (Locatello et al., 2020b). Results from other baselines are given in Appendix G, including S-TCVAE, S-FactorVAE which essentially make no difference due to the independence assumption, and the unidirectional generative model CausalGAN. In addition, we extend GraphVAE (He et al., 2018) to a supervised version, named S-GraphVAE by adding the supervised loss in the same way as DEAR and assuming the super-graph of the true graph is known a priori. However, in contrast to the composite prior in DEAR, GraphVAE assigns an SCM over the whole latent space and hence only allows a sufficiently low dimensional latent space. This makes the GraphVAE model less expressive and difficult to be applied to complex data sets with a large number of generative factors like CelebA. The qualitative results of S-GraphVAE in controllable generation are given in Appendix G. Note that we do not compare with unsupervised disentanglement methods (e.g., unsupervised β-VAE, GraphVAE, etc.) because of fairness and their lack of justification.

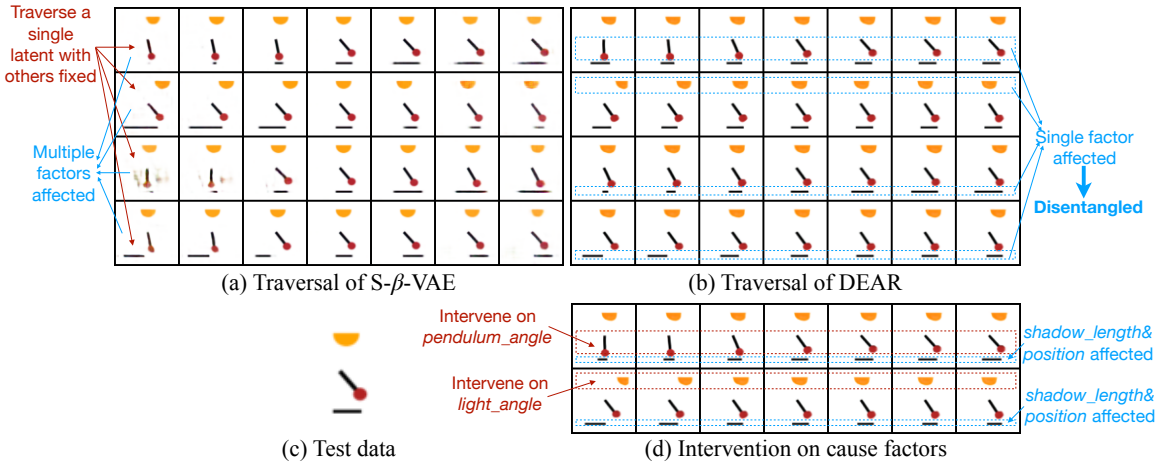


Figure 3: Results in causal controllable generation on Pendulum. For example, in line 1 of (a,b) when changing the first dimension $[z]_1$ of z which is supervised with the annotated label of *pendulum_angle* while keeping the others fixed, we see that the traversals of DEAR vary only in *pendulum_angle* (disentanglement), while those of S-β-VAE vary in both *pendulum_angle* and *shadow_length* (entanglement); in line 3 when changing $[z]_3$ with the others fixed, only *shadow_length* is affected with DEAR but both *shadow_length* and *pendulum_angle* are affected with S-β-VAE. In line 1 of (d) we see the intervening on *pendulum_angle* affects its effects *shadow_length* and *shadow_position*, which is consistent with the desired interventional distribution.

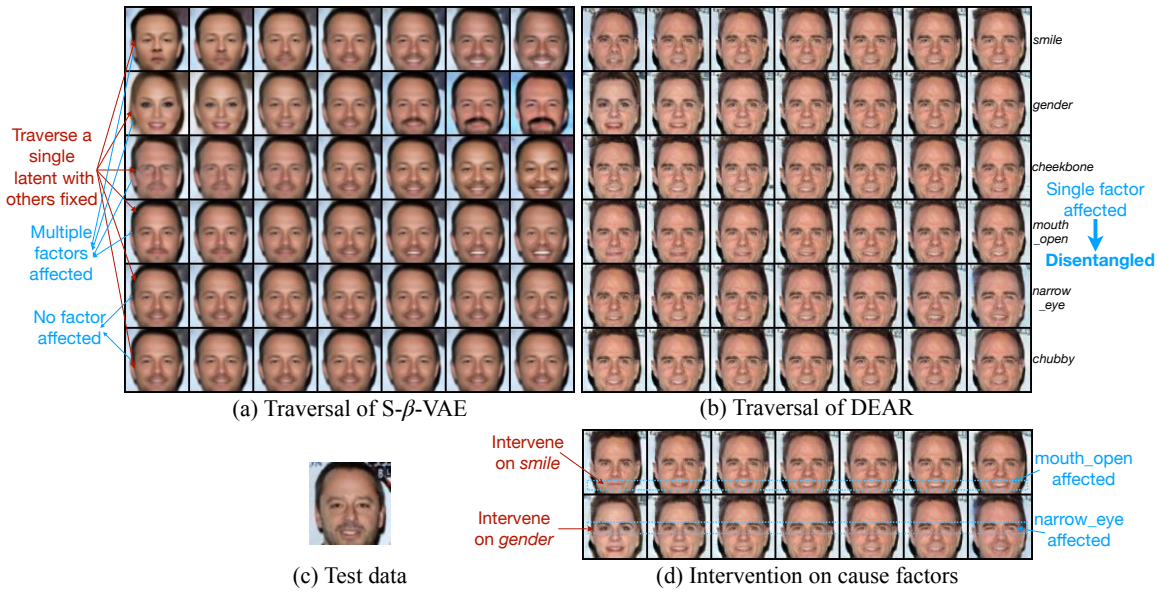


Figure 4: Results in causal controllable generation on CelebA. For example, in line 1 of (a,b) when altering $[z]_1$ with the others fixed, we see that the traversals of DEAR vary only in a single factor *smile* with factor *mouth_open* unaffected, while S-β-VAE entangles the two factors. In line 5-6 of (a), when changing $[z]_5$ and $[z]_6$ which are supervised with *narrow_eye* and *chubby*, no factors seem to be affected, indicating that the S-β-VAE fails to learn the representations of some factors. In line 1 of (d) we see that intervening on *smile* affects its effect *mouth_open*, which makes sense.

In each figure, we first infer the latent representations from a test image in block (c). The traditional traversals of the two models are given in blocks (a,b). We see that in each line when manipulating one latent dimension while keeping the others fixed, the generated images of our model vary only in a single factor, indicating that our method can disentangle the causally related factors, while those of S-β-VAE show multiple factors affected. It is worth pointing out that we are the first to achieve the disentanglement between a cause factor and its effects, while other methods tend to entangle them. One typical example is the disentanglement between *smile* and its effect *mouth_open* as shown in Figure 4. In block (d), we show the results of intervention on the latent variables representing the cause factors, which clearly show that intervening on a cause variable changes its effect variables. Results in Appendix G further show that intervening on an effect variable does not influence its cause. Specific examples are given in the captions. Note that without an SCM prior, S-β-VAE cannot generate data from general interventional distributions. More qualitative traversals from DEAR are given in Appendix G.

5.2 Downstream task

The previous section verifies the good disentanglement performance of DEAR. In this section, equipped with DEAR, we investigate and demonstrate the benefits of the learned disentangled causal representations for downstream tasks in terms of sample efficiency and distributional robustness. In Appendix B, we propose a quantitative metric for causal disentanglement which is utilized to provide some justifications on the relationship between causal disentanglement and performance in downstream tasks.

We now introduce the downstream prediction tasks. On CelebA, we consider the structure CelebA-Attractive in Figure 2(c). We artificially create a target label $\tau = 1$ if *young*=1, *gender*=0, *receding_hairline*=0, *make_up*=1, *chubby*=0, *eye_bag*=0, and $\tau = 0$ otherwise, indicating one kind of attractiveness as a slim young woman with makeup and thick hair.[3] On the pendulum data set, we regard the label of data corruption as the target τ, that is, $\tau = 1$ if the data is corrupted and $\tau = 0$ otherwise. We consider the downstream tasks of predicting the target label. In both cases, the generative factors of interests in Figure 2(a,c) are causally related to τ, which are the features that humans would use to do the task. Hence it is conjectured that a disentangled representation of these causal factors tends to be more data-efficient and invariant to distribution shifts.

5.2.1 Sample efficiency

For a BGM including the earlier state-of-the-art supervised disentanglement methods S-VAEs (Locatello et al., 2020b), the modified S-GraphVAE (He et al., 2018), and our proposed DEAR, we use the learned encoder to embed the training data to the latent space and train an MLP classifier on top of the representations to predict the target label. All the architectures are the same for various methods with details given in Appendix F. Without an encoder, one normally needs to train a convolutional neural network with raw images as the input. Here we adopt the ResNet50 (named ResNet in Table 1) as the baseline classifier which is the architecture of the BGM encoder. Since the disentanglement methods use addi-

3. Note that the definition of attractiveness here only refers to one kind of attractiveness, which has nothing to do with its linguistic definition.

tional supervision of the generative factors, we consider another baseline ResNet50 (named ResNet-pretrain) that is pretrained using multi-label classification to predict the factors on the same training set. Unless indicated otherwise, DEAR, S-VAEs, S-GraphVAE, and ResNet-pretrain have access to the annotated labels for all training samples, and DEAR and S-GraphVAE are given the true graph structure. We provide the detailed results when there is less supervised information on labels and the graph structure in Sections 5.4 and 5.3.

To measure the sample efficiency, we use the statistical efficiency score defined as the average test accuracy based on 100 samples divided by the average accuracy based on 10,000/all samples, following Locatello et al. (2019). Note that this metric may be misleading when a method always achieves poor accuracy with small and large training samples. Therefore, we also report the test accuracies with different training sample sizes to provide a comprehensive evaluation.

Table 1 presents the results, showing that DEAR owns the highest sample efficiency and test accuracy on both data sets. ResNet with raw data inputs has the lowest efficiency, although multi-label pretraining improves its performance to a limited extent. S-VAEs have better efficiency than the ResNet baselines but lower accuracy under the case with more training data. Since the encoders of all S-VAEs and DEAR share the same architecture, we explain the inferior performance of S-VAEs is mainly because the independent prior contradicts with the supervised loss as indicated in Proposition 4, making the learned representations entangled (as shown in the previous section) and less informative. On the Pendulum data with few underlying factors, S-GraphVAE outperforms the S-VAEs when training on a smaller sample, indicating that an SCM latent structure has advantages over the independent structure under the VAE framework. Nevertheless, even with the same amount of supervision (on both annotated labels and the same given graph structure), S-GraphVAE is still inferior to DEAR, potentially due to our better causal modeling and optimization based on a GAN algorithm. On the more complex data set CelebA, S-GraphVAE gives very poor performance, even worse than S-VAEs and ResNet.

In addition, we investigate the performance of DEAR under the semi-supervised setting where only 10% of the labels are available. We find that DEAR with fewer labels has comparable sample efficiency with that in the fully supervised setting, with a sacrifice in the accuracy that is yet still comparable to other baselines which use much more supervision. In Section 5.4, we provide ablation studies to show how DEAR behaves in terms of varying amounts of labeled samples and different choices of the regularization strength λ.

We also study knowing less prior information on the causal graph structure. In the last two lines of Table 1, DEAR-SG stands for the DEAR-LIN model trained with a given supergraph (which is not a full graph) of the true graph and DEAR-O stands for the DEAR-LIN model trained with a known causal ordering. We see that DEAR-SG leads to comparable performance as DEAR with the known graph structure, while DEAR-O is slightly worse but still competitive compared with other baseline methods. As we will show later, on Pendulum, DEAR-O can recover the true structure and the performance in downstream tasks is identical to that of DEAR given the true structure, so we skip showing the last two lines in Table 1(b). In Section 5.3, we investigate the performance in learning the SCM and in particular, the causal structure, given various amounts of prior information about

Method	(a) CelebA 100(%)	(a) CelebA 10,000(%)	(a) CelebA Eff(%)	(b) Pendulum 100(%)	(b) Pendulum all(%)	(b) Pendulum Eff(%)
ResNet	68.06±0.19	79.51±0.31	85.59±0.27	79.71±0.98	90.64±1.57	87.97±2.11
ResNet-pretrain	76.84±2.08	83.75±0.93	91.74±1.98	79.59±0.93	89.16±1.60	89.28±0.59
S-VAE	77.07±1.42	79.87±1.67	96.49±1.68	84.16±0.69	90.89±0.28	92.60±0.49
S-β-VAE	71.78±1.99	76.63±0.24	93.67±2.41	79.95±1.65	87.87±0.52	90.98±1.47
S-TCVAE	77.10±2.08	81.63±0.20	94.45±2.72	85.36±1.11	90.33±0.33	94.51±1.31
S-GraphVAE	67.87±1.19	72.09±0.51	94.14±1.14	86.08±1.61	91.90±0.53	93.65±1.29
DEAR-LIN	83.51±0.77	84.92±0.11	98.34±0.81	90.21±0.94	**93.31**±0.14	96.68±0.89
DEAR-NL	**84.44**±0.48	**85.10**±0.09	**99.23**±0.51	**90.62**±0.32	92.57±0.08	**97.93**±0.29
DEAR-LIN-10%	78.09±0.59	79.54±0.41	98.18±0.49	88.93±1.40	93.18±0.18	95.43±1.33
DEAR-NL-10%	80.30±0.24	80.87±0.12	**99.29**±0.23	87.65±0.46	91.27±0.21	96.03±0.29
DEAR-SG	83.69±0.63	84.91±0.06	98.57±0.67			
DEAR-O	82.84±0.68	84.42±0.05	98.13±0.79			

Table 1: Sample efficiency and test accuracy with different training sample sizes. DEAR-LIN and -NL denote the DEAR models with linear and nonlinear f respectively.

the true graph, where more insights are given to explain the comparable performance of DEAR-SG in downstream tasks.

5.2.2 Distributional robustness

We manipulate the training data to inject spurious correlations—misleading heuristics that work for most training examples but do not always hold (Sagawa et al., 2019)—between the target label and some spurious attributes. On CelebA, we regard *mouth_open* as the spurious factor; on Pendulum, we choose *background_color* $\in$ {blue(+), white(−)}. We manipulate the training data such that the target label is more strongly correlated with the spurious attributes. Specifically, the target label and the spurious attribute of 80% of the examples are both positive or negative, while those of 20% examples are opposite. For instance, in the manipulated training set, 80% smiling examples in CelebA have an open mouth; 80% corrupted examples in Pendulum are masked with a blue background. The test sets however do not have such correlations, that is, around half of the examples in the test sets of both CelebA and Pendulum have consistent target and spurious labels, leading to a distribution shift.

Intuitively these spurious attributes are not causally related to the target label, but normal independent and identically distributed (IID) based methods like empirical risk minimization (ERM) tend to exploit such easily learned spurious correlations in prediction, and hence face performance degradation when such correlation no longer exists during testing. In contrast, causal factors are regarded as invariant and thus more robust under such shifts.

Previous sections justify both theoretically and empirically that DEAR can learn disentangled causal representations well. We then apply those representations by training a classifier upon them to predict the target label, which is conjectured to be invariant and

	(a) CelebA		(b) Pendulum	
Method	**WorstAcc(%)**	**AvgAcc(%)**	**WorstAcc(%)**	**AvgAcc(%)**
ERM	59.12±1.78	82.12±0.26	60.48±2.73	87.40±0.89
ERM-multilabel	59.17±4.02	82.05±0.25	61.70±4.02	87.20±1.00
S-VAE	60.54±3.48	79.51±0.58	20.78±4.45	84.26±1.31
S-β-VAE	63.85±2.09	80.82±0.19	44.12±9.73	86.99±1.78
S-TCVAE	64.93±3.30	81.58±0.14	35.50±5.57	86.64±1.15
S-GraphVAE	50.51±4.43	76.01±1.73	54.42±4.15	87.64±2.06
DEAR-LIN	76.05±0.70	83.56±0.09	**75.60**±0.27	**93.58**±0.03
DEAR-NL	**76.98**±0.66	**83.60**±0.04	75.39±2.11	93.16±0.04
DEAR-LIN-10%	71.40±0.47	81.04±0.14	74.05±1.56	92.63±0.07
DEAR-NL-10%	70.44±1.02	81.94±0.31	73.93±1.98	92.72±0.03
DEAR-SG	74.95±1.14	83.56±0.25		
DEAR-O	74.00±1.47	83.45±0.32		

Table 2: Distributional robustness. The worst-case and average test accuracy.

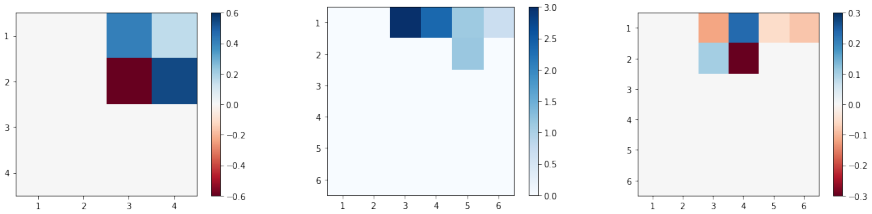
robust. Baseline methods include ERM, multi-label ERM which is trained to predict both target label and the factors considered in disentanglement in order to have the same amount of supervision, S-VAEs that are shown unable to disentangle well in the causal case, and S-GraphVAE.

Table 2 presents the average and worst-case test accuracy to assess both the overall classification performance and distributional robustness. The worst-case (Sagawa et al., 2019) accuracy refers to the following: we group the test set according to the two binary labels, the target one and the spurious attribute, into four cases and regard the group with the worst accuracy as the worst-case, which usually owns the opposite spurious correlation to the training data. It can be seen that the classifiers trained upon DEAR representations significantly outperform the baselines in both metrics. Particularly, when comparing the worst-case accuracy with the average one, we observe a slump from around 80 to around 60 for other methods on CelebA, while DEAR enjoys a much smaller decline. As in sample efficiency, S-GraphVAE suffers from a smaller drop in worst-case accuracy than S-VAEs on Pendulum, but remains inferior to DEAR. On CelebA, S-GraphVAE again shows poor performance.

Moreover, with fewer annotated samples (i.e., 10% of the full sample), DEAR-10% remains competitive against baseline methods which use even more supervised labels. DEAR-SG (given the super-graph) is slightly better than DEAR-O (given the ordering), both of which are comparable to DEAR given the full structure. More ablation studies in terms of the labeled proportion as well as the strength of the supervised regularizer are given in Section 5.4.

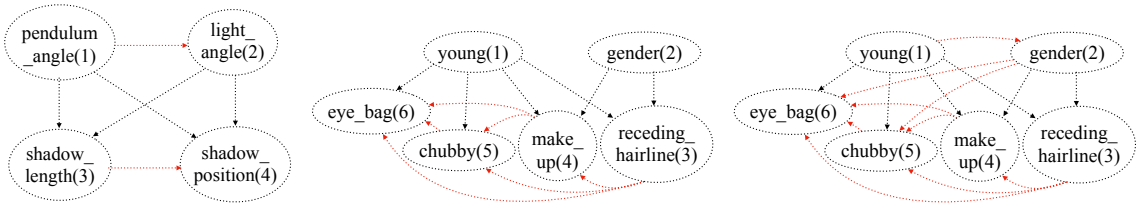
5.3 Learning of the structure A

In this section, we take a closer look into the learned causal structure and weighted adjacency matrix A of the SCM prior given various amounts of prior graph information. As mentioned in Section 4.1.2, the DEAR method requires prior knowledge on the super-graph of the true



(a) Pendulum (b) CelebA-Smile (c) CelebA-Attractive

Figure 5: The weighted adjacency matrices learned by DEAR.



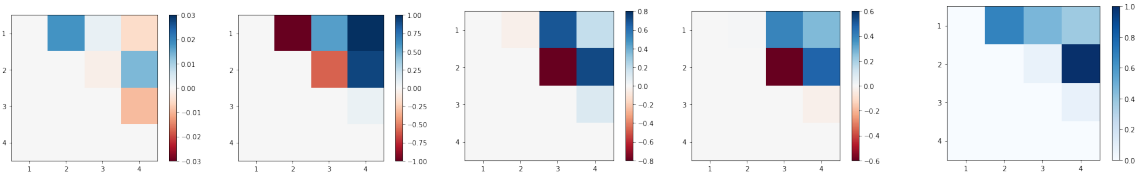
(a) Pendulum-O (b) CelebA-Attractive-SG (c) CelebA-Attractive-O

Figure 6: The given causal structures. -O and -SG stand for the causal ordering and super-graph. The black edges are true and red edges are in fact redundant.

graph over the underlying factors of interests. The experiments shown in previous sections are all based on the given true binary structure $\mathbf{I}_{A_0}$. Here we investigate the performance in learning the causal structure on knowing various amounts of information about the graph, which ranges from the causal ordering to the true structure. Note that the adjacency matrices learned by DEAR-LIN and DEAR-NL are consistent up to some scaling, so in this section we only show the results from DEAR-LIN as a representative.

Figure 5 shows the learned weighted adjacency matrices when the true binary structure is given for the three underlying structures shown in Figure 2. It can be seen that the weights exhibit meaningful signs and scalings that are consistent with common knowledge. For example, the factor *smile* and its effect *mouth_open* are positively correlated, that is, one is more likely to open mouth when smiling. The corresponding element in the weighted adjacency A_{14} of (b) turns out positive, which makes sense. Also *gender* (the indicator of being male) and its effect *make_up* are negatively correlated, that is, women tend to make up more often than men. Correspondingly, element A_{24} of (c) turns out negative.

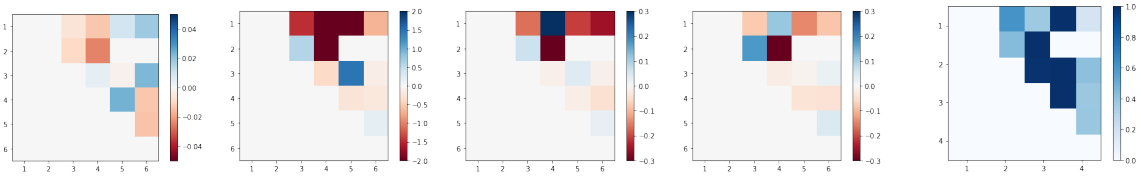
Next, we evaluate the performance of DEAR in structure learning with less prior knowledge on the true graph, i.e. knowing a super-graph rather than the exact true graph. We first study on the synthetic data set Pendulum whose ground-truth structure is shown in Figure 2(a), where there are fewer causal factors and no hidden confounder. Consider the causal ordering *pendulum_angle*, *light_angle*, *shadow_position*, *shadow_length*, given which we start with a full graph (shown in Figure 6(a)) represented by an upper triangular adjacency matrix whose elements are randomly initialized around 0 (shown in Figure 7(a)). Figure 7(a-d) present the weighted adjacency matrices learned by DEAR at different training epochs. We observe that the weights of the two redundant edges A_{12} and A_{34} vanish



(a) Epoch 0 (b) Epoch 100 (c) Epoch 200 (d) Epoch 500 (e) S-GraphVAE

Figure 7: Learned weighted adjacency matrices on Pendulum given the causal ordering. (a-d) are the learned matrices from DEAR at different training epochs starting from random initialization around 0, and (e) is the result from S-GraphVAE.

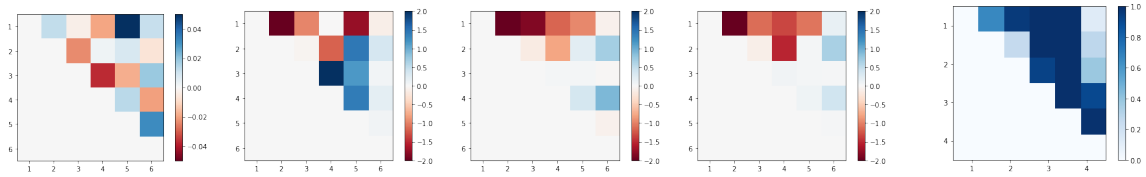
gradually and it eventually leads to the weighted adjacency that nearly coincides with the one learned given the true graph shown in Figure 5(a). In contrast, Figure 7(e) shows the structure learned by S-GraphVAE. Note that GraphVAE learns a binary structure with 0-1 elements and (e) shows the learned probabilities of each element being 1. We see that it learns a redundant edge A_{12} from *pendulum_angle* to *light_angle* and misses the edge A_{23} from *light_angle* to *shadow_position*. This experiment shows the advantage of DEAR over GraphVAE in learning the latent causal structure.



(a) Epoch 0 (b) Epoch 5 (c) Epoch 50 (d) Epoch 150 (e) S-GraphVAE

Figure 8: Learned weighted adjacency matrices on CelebA given a super-graph. (a) represents a random initialization around 0 of the weighted adjacency matrix corresponding to the super-graph in Figure 6(b); (b-d) are the learned matrices by DEAR at different training epochs; (e) is the result from S-GraphVAE.

The case is more complicated on the real data set CelebA. Although the number of factors of interests, six, is not large, there are much more underlying generative factors. Some of the other factors that we are not interested to disentangle could serve as the hidden confounders of the factors that we are interested in. For example, staying up late may cause a person to have eye bags and look chubby and hence serves as a hidden confounder of the two factors *eye_bag* and *chubby* in Figure 2(c). These hidden confounders can be captured in the remaining dimensions of the learned representations through the composite prior introduced in Section 4.1.4. However, their existence makes it difficult to identify and learn the structure of the factors of interest. Another complication comes from some biases in the data, potentially caused by selection bias or unknown interventions. Such biases may result in spurious correlations even among the causal variables, bringing trouble to causal structure learning. There are orthogonal works (e.g., Ke et al., 2019; Bengio et al., 2020;



(a) Epoch 0 (b) Epoch 5 (c) Epoch 50 (d) Epoch 150 (e) S-GraphVAE

Figure 9: Learned weighted adjacency matrices on CelebA given the causal ordering. (a-d) are the learned matrices by DEAR at different training epochs starting from random initialization around 0; (e) is the result from S-GraphVAE.

Brouillard et al., 2020) focusing on causal discovery under hidden confounders or unknown interventions, which however is beyond the scope of this paper and will be systematically explored in future work. Here we only provide some empirical studies to evaluate our method under this complicated case.

We conduct two experiments on CelebA. In the first one, we assume knowing a super-graph (Figure 6(b)) of the true graph (Figure 2(c)) and randomly initialize its weighted adjacency matrix around 0 as in Figure 8(a). Then Figure 8(a-d) show the weighted adjacency matrices learned by DEAR at different training epochs. Similar to the previous experiment on Pendulum, the weights corresponding to the redundant edges gradually vanish. Eventually, DEAR learns the weighted adjacency matrix that largely agrees with the one learned given the true graph shown in Figure 5(c). After edge pruning, one can essentially recover the true graph structure. This explains why DEAR-SG (the DEAR model given this super-graph) performs competitively with DEAR given the true structure in the downstream tasks in the previous two sections. In contrast, the graph learned by GraphVAE shown in Figure 8(e) fails to recover the true structure, although it is given the same known super-graph as DEAR.

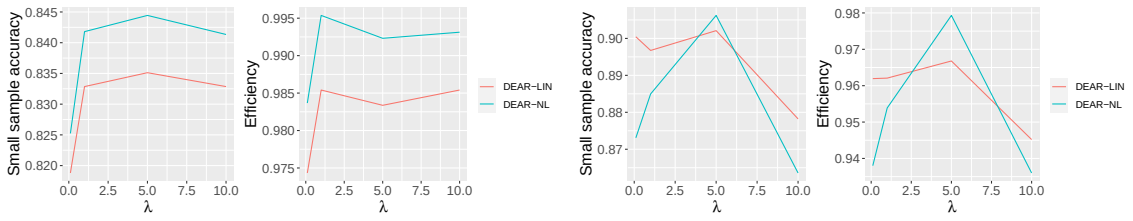
In the second experiment, we only assume knowing the causal ordering which leads to a full graph shown in Figure 6(c) with the upper-triangular weighted adjacency matrix randomly initialized in Figure 9(a). We observe that although DEAR can remove most of the redundant edges, it mistakenly learns a large weight on the edge from *young* to *gender*. This may be due to the spurious correlation between the two factors *young* and *gender* potentially caused by the selection bias during data collection. In comparison, as shown in Figure 9(e), the graph learned by GraphVAE given the same causal ordering turns out to be farther away from the true graph than DEAR. Nevertheless, as discussed in the previous two sections, DEAR-O (the DEAR model given the causal ordering) still achieves reasonably satisfying performance, which indicates the robustness of our DEAR method against the correctness of the learned graph structure.

In summary, when given the true graph structure, DEAR can learn meaningful weights for each edge. If there is no hidden confounder or spurious correlation among the factors of interests, DEAR can learn the true graph given only the causal ordering. If there exist such biases, DEAR can only recover the true structure given some proper super-graphs and in general cannot learn all edges correctly when only the causal ordering is given. In all cases, DEAR outperforms GraphVAE in learning the causal structure.

5.4 Ablation study

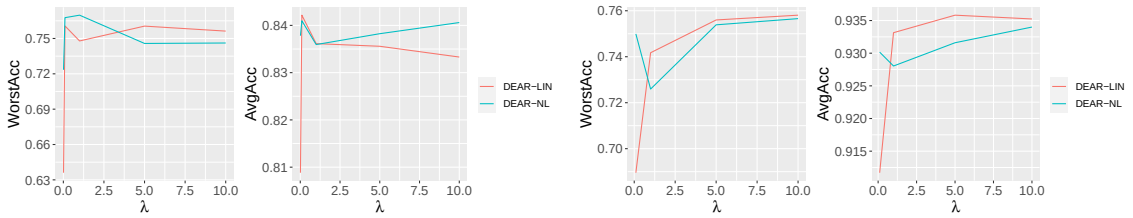
In this section, we conduct ablation studies to illustrate how DEAR performs when using different choices of the hyperparameter λ which determines the weight of the supervised regularizer and varying amounts of labeled samples. According to Proposition 5 and Theorem 8, at the population level, i.e., assuming an infinite amount of data, the regularization strength λ in the objective (7) can be any arbitrary positive value to make the theorems hold. However, in practice with a finite sample, λ cannot be arbitrarily small roughly due to the estimation error. Therefore we suggest regarding λ as a hyperparameter and investigate its sensitivity across different tasks and data sets. Figures 10-11 plot the metrics in sample efficiency and distributional robustness when using different choices of λ. We observe that all these results (with λ ranging from 0.1 to 10) remain significantly superior to the baseline methods in Tables 1-2, which suggests that DEAR can perform reasonably well across a wide range of λ. As λ becomes close to 0, we generally observe a performance decrease.

Next, we study how DEAR, as well as baseline methods, behave as we reduce the number of annotated samples. Figures 12-13 plot the metrics in sample efficiency and distributional robustness when using different amounts of labeled samples. Note that 0.1% of the CelebA training set corresponds to 162 samples and 1% of the Pendulum training set corresponds to 67 samples, both of which belong to weakly supervised settings according to Locatello et al. (2020b). Such small numbers of supervised labels belong to weakly supervised settings according to Locatello et al. (2020b) and would make manual labeling feasible even if no label is available beforehand. Naturally, with fewer labeled samples, all methods basically perform worse. DEAR always outperforms the VAEs. In particular, as shown in Figure 13(a), when training with 0.1%-1% labels of the CelebA training sample, S-β-VAE and S-TCVAE completely fail in the worst-case group, meaning that the classifiers trained upon them almost fully rely on the spurious correlation and exhibit no robustness to distribution shifts at all. In Figure 12(a), when the supervised proportion is lower, although S-β-VAE and S-TCVAE have higher sample efficiency, they actually perform poorly with both small and large samples, leading to a misleadingly high efficiency score.



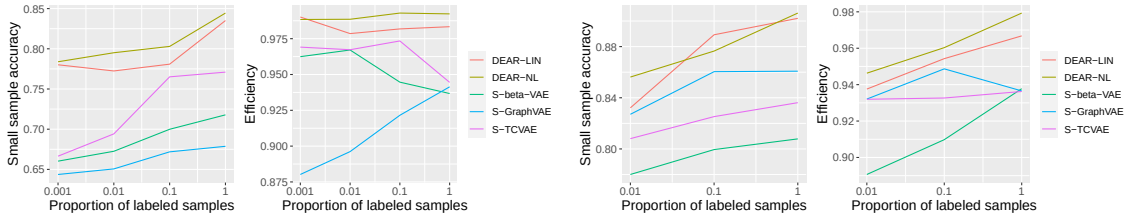
(a) CelebA (b) Pendulum

Figure 10: Test accuracy when training on a small sample & sample efficiency, as defined in Section 5.2.1, against four different choices of λ: 0.1, 1, 5, and 10.



(a) CelebA (b) Pendulum

Figure 11: Worst-case and average test accuracy, as defined in Section 5.2.2, against different choices of λ. On Pendulum, we experiment with $\lambda = 0.1, 1, 5, 10$; on CelebA, we experiment with $\lambda = 0.01, 0.1, 1, 5, 10$.



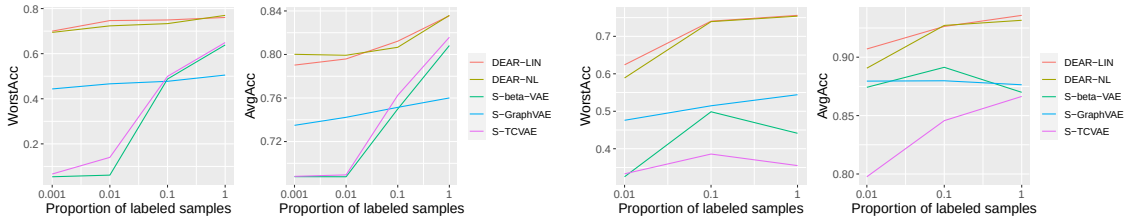
(a) CelebA (b) Pendulum

Figure 12: Test accuracy with a small training sample & sample efficiency against different proportions of labeled samples among full data. On the larger data set CelebA, we consider proportion=0.001, 0.01, 0.1, 1; on the smaller Pendulum data, we consider 0.01, 0.1, 1.

6. Conclusion

In this paper, we showed that previous methods with the independent latent prior assumption fail to learn disentangled representation when the underlying factors of interests are causally related. We then proposed a new disentangled learning method called DEAR with theoretical guarantees for identifiability and asymptotic consistency. Extensive experiments demonstrated the effectiveness of DEAR in causal controllable generation and structure learning, and the benefits of the learned representations for downstream tasks.

Several future directions are worth exploring. Although in our ablation experiments, we demonstrated that DEAR exhibits promising performance in weakly supervised settings in terms of annotated labels and the graph structure, it is worth considering more flexible forms of supervision to make DEAR widely adopted in more real-world applications. On one hand, regarding the annotated labels of the factors of interests, one may consider utilizing other forms of supervision, such as restricted labeling or rank pairing (Shu et al., 2020). Besides, instead of using direct supervision about the true factors, one may consider some additionally observed variables such as class labels or time index (Khemakhem et al., 2020) which serve as auxiliary information to ensure more general identifiability of the true latent factors in the causal case. On the other hand, regarding the graph structure, our experiments in Section 5.3 indicated the potential of DEAR in latent structure learning. As in many real applications, even the causal ordering may not be available, it is promising



(a) CelebA (b) Pendulum

Figure 13: Worst-case and average test accuracy against different proportions of labeled samples among full data.

to incorporate causal discovery methods in the DEAR framework to learn the latent causal structure from scratch (i.e., without any prior information) with a guarantee of the structure identifiability.

In addition, the proposed method applies to the case where the observational data are IID, as commonly considered in the literature of generative models and disentanglement. It would be interesting to extend the current approach to non-IID settings, in particular, to the scenarios where one can perform interventions during data collection. For example, in reinforcement learning, the interactive environment allows the agent to perform actions and observe their outcomes. The resulting data set that contains a mixture of interventional distributions (e.g., Ke et al., 2021) could be leveraged in causal disentanglement learning.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments that were very useful for improving the quality of this work. The work was supported by the General Research Fund (GRF) of Hong Kong (No. 16201320). F. Liu's research was supported in part by a Key Research Project of Zhejiang Lab (No. 2022PE0AC04).

Appendix A. Proofs

A.1 Preliminaries

This section presents some preliminary notions and lemmas which will be used in proofs.

Definition 9 (Bracketing covering number (van de Geer, 2000)) *Consider a function class $\mathcal{G} = \{g(x)\}$ and a probability measure μ defined on $\mathcal{X}$. Given any positive number $\delta > 0$. Let $N_{1,B}(\delta, \mathcal{G}, \mu)$ be the smallest value of N for which there exist pairs of functions $\{[g_j^L, g_j^U]\}_{j=1}^N$ such that $\int |g_j^L(x) - g_j^U(x)| d\mu \leq \delta$ for all $j = 1, \ldots, N$, and such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$ such that $g_j^L \leq g \leq g_j^U$. Then $N_{1,B}(\delta, \mathcal{G}, \mu)$ is called the δ-bracketing covering number of $\mathcal{G}$.*

Lemma 10 (Uniform continuous mapping theorem) *Let X_n, X be random vectors defined on $\mathcal{X}$. Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be uniformly continuous and $T_\theta : \mathcal{X} \to \mathbb{R}^d$ for $\theta \in \Theta$. Suppose $T_\theta(X_n)$ converges uniformly in probability to $T_\theta(X)$ over Θ, i.e., as $n \to \infty$ we have $\sup_{\theta \in \Theta} \|T_\theta(X_n) - T_\theta(X)\| \xrightarrow{p} 0$. Then $f(T_\theta(X_n))$ converges uniformly in probability to $f(T_\theta(X))$, i.e., $\sup_\theta \|f(T_\theta(X_n)) - f(T_\theta(X))\| \xrightarrow{p} 0$.*

Proof Given any $\epsilon > 0$. Because f is uniformly continuous, there exists $\delta > 0$ such that $\|f(x) - f(y)\| \leq \epsilon$ for all $\|x - y\| \leq \delta$.

We have

$$\begin{aligned} \mathbb{P}\Big(\sup_{\theta \in \Theta} \|T_\theta(X_n) - T_\theta(X)\| \leq \delta\Big) &= \mathbb{P}\big(\forall \theta \in \Theta : \|T_\theta(X_n) - T_\theta(X)\| \leq \delta\big) \qquad (11) \\ &\leq \mathbb{P}\big(\forall \theta \in \Theta : \|f(T_\theta(X_n)) - f(T_\theta(X))\| \leq \epsilon\big) \\ &= \mathbb{P}\Big(\sup_{\theta \in \Theta} \|f(T_\theta(X_n)) - f(T_\theta(X))\| \leq \epsilon\Big). \qquad (12) \end{aligned}$$

By the uniform convergence of $T_\theta(X_n)$, we know the left-hand side of (11) converges to 1. Hence (12) goes to 1, which implies the desired result. ■

Lemma 11 *Let μ_n and μ be a sequence of measures on probability space $(\mathcal{X}, \Sigma)$ with densities $p_n(x)$ and $p(x)$. Given any compact subset K of $\mathcal{X}$. Suppose p_n is uniformly bounded and Lipschitz on K $(*)$. If $H^2(\mu_n, \mu) \xrightarrow{p} 0$, then $\sup_{x \in K} |p_n(x) - p(x)| \xrightarrow{p} 0$ as $n \to \infty$, where $H(q_1, q_2) = \Big(\int (q_1^{1/2} - q_2^{1/2}) dx dz/2\Big)^{1/2}$ denotes the Hellinger distance between two distributions with densities q_1 and q_2.*

Proof Note that assumptions in $(*)$ satisfy the requirements in the Arzelà-Ascoli theorem. Thus, for each subsequence of p_n, there is a further subsequence p_{n_m} which converges uniformly on compact set K, i.e., for some p_0 as $m \to \infty$ we have

$$\sup_{x \in K} |p_{n_m}(x) - p_0(x)| \to 0.$$

By Scheffé's Theorem we have $H(p_{n_m}, p_0) \to 0$. On the other hand we have $H(p_{n_m}, p) \xrightarrow{p} 0$. By triangle inequality,

$$H(p, p_0) \leq H(p_{n_m}, p_0) + H(p_{n_m}, p) \xrightarrow{p} 0.$$

Since the inequality holds for all m and the LHS is deterministic, we have $H(p, p_0) = 0$, which implies $p = p_0$, a.e. wrt the Lebesgue measure. Hence we have

$$\sup_{x \in K} |p_{n_m}(x) - p(x)| \to 0, \ a.e.$$

By Durrett (2019, Theorem 2.3.2), we have $\sup_{x \in K} |p_n(x) - p(x)| \xrightarrow{p} 0$ as $n \to \infty$. ■

A.2 Proof of Proposition 4

Proof On one hand, by the assumption that the elements of ξ are connected by a causal graph whose adjacency matrix is not a zero matrix, there exist $i \neq j$ such that $[\xi]_i$ and $[\xi]_j$ are not independent, indicating that the probability density of ξ cannot be factorized. Since E^* is disentangled with respect to ξ, by Definition 3, $\forall i = 1, \ldots, m$ there exists g_i such that $[E^*(x)]_i = g_i([\xi]_i)$. This implies that the probability density of $E^*(x)$ is not factorized.

On the other hand, notice that the distribution family of the latent prior is contained in $\{p_z : p_z \text{ is factorized}\}$. Hence the intersection of the marginal distribution families of z and $E^*(x)$ is an empty set. Then the joint distribution families of $(x, E^*(x))$ and $(G(z), z)$ also have an empty intersection.

We know that $L_{\text{gen}}(E^*, G) = 0$ implies $q_{E^*}(x, z) = p_G(x, z)$ which contradicts the above. Therefore, we have $a = \min_G L_{\text{gen}}(E^*, G) > 0$.

Let (E', G') be the solution of the optimization problem $\min_{\{(E,G): L_{\text{gen}}=0\}} L_{\text{sup}}(E)$. From the above we know E' cannot be disentangled with respect to ξ. Then we have $L' = L(E', G') = \lambda b$, and $L^* = L(E^*, G) \geq a + \lambda b^* > \lambda b^*$ for any generator G. When $b^* \geq b$ we directly have $L' < L^*$. When $b^* < b$ and λ is not large enough, i.e., $\lambda < \frac{a}{b - b^*}$, we have $L' < L^*$. ■

Discussion on Träuble et al. (2021, Proposition 1)

Proposition 1 in Träuble et al. (2021) and our Proposition 4 state the same unidentifiability issue from different perspectives. Proposition 1 in Träuble et al. (2021) says that maximum likelihood estimation (MLE) cannot identify the disentangled representation, while our Proposition 4 says that the formulation (7) in our paper cannot identify the disentangled representation. The relationship of the two formulations, MLE and (7), is that the first term in (7) is an upper bound of the negative log-likelihood. Therefore, our Proposition 4 is more straightforward in the sense that it directly studies the formulation that is used in disentanglement methods.

A.3 Proof of Proposition 5

In this section, we prove a full statement of Proposition 5. Specifically, we add an assumption on structure identifiability and the consequent result in learning the true structure. Assumption 2 states the identifiability of the true causal structure $\mathbf{I}_{A_0}$ of ξ, which

is applicable given the true causal ordering under the basic Markov and causal minimality conditions (Pearl, 2014; Zhang and Spirtes, 2011).

Assumption 2 *For all $\beta = (f, h, A) \in \mathcal{B}$ with $p_\beta = p_{\beta_0}$, it holds that $\mathbf{I}_A = \mathbf{I}_{A_0}$.*

Proposition 12 (Full statement of Proposition 5) *Assume the infinite capacity of E and G. Further under Assumptions 1 and 2, DEAR formulation (7) learns the disentangled encoder E^* and the true causal structure $\mathbf{I}_{A_0}$. Specifically, we have $g_i(x) = \sigma^{-1}(x)$ with the CE loss as the supervised regularizer, and $g_i(x) = x$ with the L_2 loss.*

Proof To simplify the notations in this section, for a vector x, let x_i denote the i-th element of x instead of $[x]_i$. For a vector function $g(x)$, let $g_i(x)$ denote the i-th component function.

Assume E is deterministic.

On one hand, for each $i = 1, \ldots, m$, first consider the cross-entropy loss

$$\begin{aligned} L_{\text{sup},i}(E) &= \mathbb{E}_{(x,y)}[\text{CE}(E_i(x), y_i)] \\ &= -\int q_x(x)p(y_i|x)[y_i \log \sigma(E_i(x)) + (1 - y_i)\log(1 - \sigma(E_i(x)))]dxdy_i, \end{aligned}$$

where $p(y_i|x)$ is the probability mass function of the binary label y_i given x, characterized by $\mathbb{P}(y_i = 1|x) = \mathbb{E}(y_i|x)$ and $\mathbb{P}(y_i = 0|x) = 1 - \mathbb{E}(y_i|x)$. Let

$$\frac{\partial L_{\text{sup},i}}{\partial \sigma(E_i(x))} = \int q_x(x)p(y_i|x)\left(\frac{1}{1 - \sigma(E_i)} - y_i\frac{1}{\sigma(E_i)(1 - \sigma(E_i))}\right)dxdy_i = 0.$$

Then we know that $E_i^*(x) = \sigma^{-1}(\mathbb{E}(y_i|x)) = \sigma^{-1}(\xi_i)$ minimizes $L_{\text{sup},i}$.

Consider the L_2 loss

$$L_{\text{sup},i}(\phi) = \mathbb{E}_{(x,y)}[E_i(x) - y_i]^2 = \int q_x(x)p(y_i|x)[E_i(x) - y_i]^2 dxdy_i.$$

Let

$$\frac{\partial L_{\text{sup},i}}{\partial E_i(x)} = 2\int q_x(x)p(y_i|x)(E_i(x) - y_i)dxdy_i = 0.$$

Then we know that $E_i^*(x) = \mathbb{E}(y_i|x) = \xi_i$ minimizes $L_{\text{sup},i}$ in this case.

On the other hand, by Assumption 1 there exists $\beta_0 = (f_0, h_0, A_0)$ such that $p_\xi = p_{\beta_0}$. Further due to the infinite capacity of G and Assumption 1, we have the distribution family of $p_{G,F}(x, z)$ contains $q_{E^*}(x, z)$. Then by minimizing the loss in (7) over G, we can find G^* and F^* such that $p_{G^*,F^*}(x, z)$ matches $q_{E^*}(x, z)$ and thus $L_{\text{gen}}(E^*, G^*, F^*)$ reaches 0, where F^* corresponds to parameter $\beta^* = (f^*, h^*, A^*)$.

Note that $p_{G^*,F^*}(x, z) = q_{E^*}(x, z)$ implies that the marginal distributions match, i.e., $p_{F^*}(z) = q_{E^*}(z)$. Generally denote $E_i^*(x) = g_i(\xi_i)$ for $i = 1, \ldots, m$. Then, for $i = 1, \ldots, m$, the distributions of $g_i^{-1}(E_i^*(x)) = \xi_i$ and $g_i^{-1}(F_i^*(\epsilon))$ are identical. It can be seen that $p_{\beta_0} = p_{\beta_0^*}$ with $\beta_0^* = (g^{-1} \circ f^*, h^*, A^*)$, where $\circ$ denotes elementwise composition. Then according to Assumption 2, we have $\mathbf{I}_{A^*} = \mathbf{I}_{A_0}$.

Hence minimizing $L = L_{\text{gen}} + \lambda L_{\text{sup}}$, which is the DEAR formulation (7), leads to the solution with $E_i^*(x) = g_i(\xi_i)$ with $g_i(\xi_i) = \sigma^{-1}(\xi_i)$ if CE loss is used, and $g_i(\xi_i) = \xi_i$ if L_2 loss is used, and the true binary adjacency matrix $\mathbf{I}_{A_0}$.

For a stochastic encoder, we establish the disentanglement of its deterministic part as above, and follow Definition 3 to obtain the desired result. ■

A.4 Proof of Lemma 7

Proof [Proof of Lemma 7] The proof proceeds in three steps after an introduction on logistic regression under the scenario of generative models.

For pair (x, z), let label $w = 1$ if $(x, z) \sim q_E$ and $w = 0$ if $(x, z) \sim p_G$, which states that $p(x, z|w = 1) = q_E(x, z)$ and $p(x, z|w = 0) = p_{G,F}(x, z)$. In generative models, the prior is given by $\mathbb{P}(w = 1) = \mathbb{P}(w = 0) = 1/2$. Then the marginal distribution of (x, z) is given by $p^*(x, z) = q_E(x, z)/2 + p_{G,F}(x, z)/2$ which induces the probability measure μ^*. Note that the analysis below holds for all $\boldsymbol{\theta} \in \Theta$ in a pointwise manner unless indicated otherwise, so for simplicity of notation, we omit the subscript $\boldsymbol{\theta}$. By the Bayes formula we have $\mathbb{P}(w = 1|x, z) = q_E(x, z)/(q_E(x, z) + p_{G,F}(x, z))$ and $\mathbb{P}(w = 0|x, z) = p_{G,F}(x, z)/(q_E(x, z) + p_{G,F}(x, z))$ which defines the probability mass function $p^*(w|x, z)$. Let $p^*(x, z, w) = p^*(x, z)p^*(w|x, z)$.

Recall the definition $D^*(x, z) = \log(q_E(x, z)/p_{G,F}(x, z))$, so we notice $\mathbb{P}(w = 1|x, z) = 1/(1 + e^{-D^*(x,z)})$. Consider the family of conditional distributions $\mathcal{P} = \{P_D(w = 1|x, z) = 1/(1 + e^{-D(x,z)}) : D \in \mathcal{D}\}$.

Logistic regression maximizes the log-likelihood

$$\mathbb{E}_{x,z,w\sim p^*(x,z,w)}[\log p_D(x, z, w)] = \mathbb{E}_{p^*(x,z,w)} \log[p^*(x, z)p_D(w|x, z)]$$

over $\mathcal{P}$ or equivalently over $\mathcal{D}$.

Given IID samples $(x_i, z_i, w_i), i = 1, \ldots, N_d$ from $p^*(x, z, w)$, the empirical loss to be minimized is given by

$$\begin{aligned}\hat{L}_d(D) &= -\frac{1}{N_d}\sum_{i=1}^{N_d}[\log p_D(w_i|x_i, z_i) + \log p^*(x_i, z_i)] \\ &= \frac{1}{N_d}\Bigg[\sum_{i:w_i=1} \log(1 + e^{-D(x_i,z_i)}) + \sum_{i:w_i=0} \log(1 + e^{D(x_i,z_i)})\Bigg] + c,\end{aligned}$$

which is equivalent to (9) up to a constant c. Let $\hat{D} = \operatorname{argmin}_{D\in\mathcal{D}} \hat{L}_d(D)$ and $\hat{P}(w = 1|x, z) = 1/(1 + e^{-\hat{D}(x,z)})$.

Step I We now establish the consistency of $\hat{D}(x, z)$ to $D^*(x, z)$ as defined in (14) below based on the generalization analysis of maximum likelihood estimation.

Let the class

$$\mathcal{G} = \left\{g(x, z, w) = \frac{1}{2}\log\frac{p_D(x, z, w) + p^*(x, z, w)}{2p^*(x, z, w)} : D \in \mathcal{D}\right\}.$$

Note that each element of $\mathcal{G}$ can be written as

$$g(x, z, w) = \frac{1}{2}\log\frac{p_D(w|x, z) + p^*(w|x, z)}{2p^*(w|x, z)}.$$

By the boundedness of D^* in condition *C4*, we know the mass function $p^*(w|x,z)$ is bounded in a closed interval within $(0,1)$, so g is uniformly bounded. Let $g_\infty = \sup_{g\in\mathcal{G}}|g|$. Then $\mathbb{E}_{p^*(x,z,w)}[g_\infty(x,z,w)] < \infty$. Moreover for all $\delta > 0$, the compactness of $\mathcal{D}$ assumed in condition *C2* implies a finite bracketing covering number defined in Definition 9, i.e., $N_{1,B}(\delta, \mathcal{D}, \mu^*) < \infty$. Then it follows from van de Geer (2000, Theorem 4.3) that

$$H(p_{\hat{D}}(x,z,w), p^*(x,z,w)) \to 0 \tag{13}$$

almost surely as $N_d \to \infty$, where H denotes the Hellinger distance.

Consider any compact subset K of $\mathcal{X}\times\mathcal{Z}$. We know that for all $D \in \mathcal{D}$, $D(x,z)$ is continuous and thus is bounded and Lipschitz on K. Also from the boundedness of D^*, we know that $p^*(x,z)$ is bounded away from 0 on K. Then it follows from (13) and Lemma 11 that

$$\sup_{(x,z)\in K} |\hat{P}(w=1|x,z) - \mathbb{P}(w=1|x,z)| \xrightarrow{p} 0.$$

Then by continuous mapping theorem (Lemma 10) and noting that $l(p) = \log(p/(1-p))$ is uniformly continuous on a closed interval within $(0,1)$, we have as $N_d \to \infty$

$$\sup_{(x,z)\in K} |\hat{D}(x,z) - D^*(x,z)| \xrightarrow{p} 0. \tag{14}$$

Step II We then prove the pointwise consistency of $\nabla\hat{D}(x,z)$ to $\nabla D^*(x,z)$ as defined in (17).

Construct an arbitrary probability measure μ on $\mathcal{X}\times\mathcal{Z}$ that satisfies the following (e.g., a Gaussian measure):

- μ is absolutely continuous with respect to Lebesgue measure with a density ρ.
- μ is tight, i.e., for any $\epsilon > 0$, there is a compact subset K_ϵ of $\mathcal{X}\times\mathcal{Z}$ such that $\mu(K_\epsilon) \geq 1-\epsilon$.
- $\nabla\log\rho$ is uniformly bounded, i.e., there exists $B_1 > 0$ such that $\|\nabla\log\rho(x,z)\| \leq B_1$ for all $(x,z)\in\mathcal{X}\times\mathcal{Z}$.
- ρ vanishes at infinity rapidly enough, i.e., $\rho(x,z) = o(r^{-d-k})$ with $r = \sqrt{\|x\|^2+\|z\|^2}$.

For a function u that is uniformly bounded on $\mathcal{X}\times\mathcal{Z}$, we have from integration by parts and Cauchy-Schwartz inequality that

$$\begin{aligned}\int \|\nabla u\|^2 d\mu &= -\int u\ tr(\nabla^2 u)d\mu - \int u\nabla u^\top \nabla\log\rho d\mu \\ &\leq \sqrt{\int |u|^2 d\mu \int [tr(\nabla^2 u)]^2 d\mu} + \sqrt{\int |u|^2 d\mu \int (\nabla u^\top \nabla\log\rho)^2 d\mu}.\end{aligned} \tag{15}$$

Recall from condition *C4* that there exists a positive number $B_0 < \infty$ such that $\forall x, z$, $\forall D\in\mathcal{D}$, we have $|D(x,z)| \leq B_0$, $\|\nabla D(x,z)\| \leq B_0$ and $|tr(\nabla^2 D(x,z))| \leq B_0$.

Given arbitrary $\epsilon > 0$, we know from the tightness of μ that there exits a compact subset K_ϵ of $\mathcal{X} \times \mathcal{Z}$ such that $\mu(K_\epsilon) \geq 1 - \epsilon$. Let $B = \max\{B_0, B_1\}$. Then we have for all $\boldsymbol{\theta} \in \Theta$ that

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Z}} \|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\|^2 d\mu \\
&\leq 2B\sqrt{\int_{\mathcal{X}\times\mathcal{Z}} |\hat{D}(x,z) - D^*(x,z)|^2 d\mu} + \sqrt{2}B^2\sqrt{\int_{\mathcal{X}\times\mathcal{Z}} |\hat{D}(x,z) - D^*(x,z)|^2 d\mu} \\
&= (2B + \sqrt{2}B^2)\sqrt{\int_{K_\epsilon} |\hat{D}(x,z) - D^*(x,z)|^2 d\mu + \int_{K_\epsilon^c} |\hat{D}(x,z) - D^*(x,z)|^2 d\mu} \\
&\leq (2B + \sqrt{2}B^2)\sqrt{\int_{K_\epsilon} |\hat{D}(x,z) - D^*(x,z)|^2 d\mu + 2B^2\epsilon},
\end{aligned} \tag{16}$$

where $K_\epsilon^c = (\mathcal{X} \times \mathcal{Z}) \setminus K_\epsilon$ is the complement, the first inequality is an application of (15) and further due to the boundedness of $\nabla \log \rho$ and gradients and Hessians of functions in $\mathcal{D}$, and the second inequality comes from the boundedness of functions in $\mathcal{D}$ and the tightness of μ.

By the uniform convergence in (14) over K_ϵ, we have for all $(x, z) \in K_\epsilon$, there exist a sequence $a_{N_d} = o_p(1)$ which is free of (x, z) such that $|\hat{D}(x,z) - D(x,z)|^2 \leq a_{N_d}$. Then we have

$$\int_{K_\epsilon} |\hat{D}(x,z) - D^*(x,z)|^2 d\mu \leq \int_{K_\epsilon} a_{N_d} d\mu = a_{N_d}\mu(K_\epsilon) = o_p(1),$$

by noting that μ is finite. Further by the arbitrariness of ϵ, we let $\epsilon \to 0$ and obtain from (16) that

$$\int_{\mathcal{X}\times\mathcal{Z}} \|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\|^2 d\mu \xrightarrow{p} 0.$$

Recall the arbitrariness of μ. For all $(x, z) \in \mathcal{X} \times \mathcal{Z}$, construct μ such that $\rho(x, z) > 0$. Let $v(x, z) = \|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\|^2 \rho(x,z)$. By the converse of the mean value theorem, if (x, z) is not an extremum of v, then there exists a bounded subset $S(x, z) \subseteq \mathcal{X} \times \mathcal{Z}$ such that

$$\begin{aligned}
\|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\|^2 \rho(x,z) &= \frac{1}{\nu(S(x,z))} \int_{S(x,z)} v(x', z') dx' dz' \\
&\leq \frac{1}{\nu(S(x,z))} \int_{\mathcal{X}\times\mathcal{Z}} v(x', z') dx' dz' \xrightarrow{p} 0
\end{aligned}$$

where ν denotes the Lebesgue measure. Since $\rho(x, z) > 0$, this implies $\|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\| \xrightarrow{p} 0$ for all non-extrema. By Lipschitz continuity of v on any compact set, we have $\|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\| \xrightarrow{p} 0$ for all extrema.

Up to now we have shown that for all $\boldsymbol{\theta} \in \Theta$ and $(x, z) \in \mathcal{X} \times \mathcal{Z}$, we have $\|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\| \xrightarrow{p} 0$ as $N_d \to \infty$. Further from the smoothness in condition *C1* and the compactness of Θ, we have $\forall x, z$, as $N_d \to \infty$

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla \hat{D}(x,z) - \nabla D^*(x,z)\| \xrightarrow{p} 0. \tag{17}$$

Step III Based on the convergence statements established above, we proceed to show the consistency of the approximate gradient $h_{\hat{D}}(\boldsymbol{\theta})$ and complete the proof.

By condition *C3*, $\{\mu^*\}$ is uniformly tight. For arbitrary $\epsilon > 0$, there exists a compact subset K_ϵ of $\mathcal{X} \times \mathcal{Z}$ such that $\mu^*(K_\epsilon^c) < \epsilon$. Because $\nabla D(x, z)$ is Lipschitz continuous with respect to (x, z) on K_ϵ, we have as $N_d \to \infty$

$$\sup_{\boldsymbol{\theta}\in\Theta,(x,z)\in K_\epsilon} \|\nabla \hat{D}(x, z) - \nabla D^*(x, z)\| \xrightarrow{p} 0. \tag{18}$$

Given any $S_N = \{(x_i, z_i) \sim \mu^*, y_j : i = 1, \ldots, N, j = 1, \ldots, N_s\}$ and $\delta > 0$. Define events $A_{N_d} = \{\sup_{\boldsymbol{\theta}} \|h_{\hat{D}}(\boldsymbol{\theta}) - h_{D^*}(\boldsymbol{\theta})\| \leq \delta\}$ and $B_{N,\epsilon} = \{\forall i : (x_i, z_i) \in K_\epsilon\}$. We have from the tightness of μ^* that $\mathbb{P}(B_{N,\epsilon}) \geq (1 - \epsilon)^N$. We know from (18) and the continuous mapping theorem (Lemma 10) that for any S_N and $\epsilon > 0$, as $N_d \to \infty$ (free of S_N), we have $\mathbb{P}(A_{N_d}|B_{N,\epsilon}) \to 1$. Then as $N_d \to \infty$, we have

$$\mathbb{P}(A_{N_d}) \geq \mathbb{P}(A_{N_d} \cap B_{N,\epsilon}) = \mathbb{P}(A_{N_d}|B_{N,\epsilon})\mathbb{P}(B_{N,\epsilon}) \geq \mathbb{P}(A_{N_d}|B_{N,\epsilon})(1 - \epsilon)^N \to (1 - \epsilon)^N.$$

By letting $\epsilon \to 0$ we have $\mathbb{P}(A_{N_d}) \to 1$ as $N_d \to \infty$. Since δ is arbitrary, we have that for any S_N, $\sup_{\boldsymbol{\theta}\in\Theta} \|h_{\hat{D}}(\boldsymbol{\theta}) - h_{D^*}(\boldsymbol{\theta})\| \xrightarrow{p} 0$ as $N_d \to \infty$.

On the other hand, by condition *C5* and the boundedness of D^*, and according to the gradient formulas in Lemma 6, it follows from the uniform law of large numbers that $\|h_{D^*}(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta})\| \xrightarrow{p} 0$ uniformly over Θ as $N, N_s \to \infty$.

By triangle inequality we have

$$\sup_{\boldsymbol{\theta}\in\Theta} \|h_{\hat{D}}(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta})\| \leq \sup_{\boldsymbol{\theta}\in\Theta} \|h_{\hat{D}}(\boldsymbol{\theta}) - h_{D^*}(\boldsymbol{\theta})\| + \sup_{\boldsymbol{\theta}\in\Theta} \|h_{D^*}(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta})\|.$$

Therefore there exists a sequence of $(N, N_s, N_d) \to \infty$ such that

$$\sup_{\boldsymbol{\theta}\in\Theta} \|h_{\hat{D}}(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta})\| \xrightarrow{p} 0$$

which completes the proof. ■

A.5 Proof of Theorem 8

Proof [Proof of Theorem 8] Consider the gradient descent step based on the approximate gradient

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta h_{\hat{D}}(\boldsymbol{\theta}_{t-1}),$$

where η is the learning rate.

Suppose $L(\boldsymbol{\theta})$ is ℓ_0-smooth. Then we have

$$L(\boldsymbol{\theta}_t) \leq L(\boldsymbol{\theta}_{t-1}) - \eta h_{\hat{D}}(\boldsymbol{\theta}_{t-1})^\top \nabla L(\theta_{t-1}) + \frac{\eta^2 \ell_0}{2} h_{\hat{D}}(\boldsymbol{\theta}_{t-1})^\top h_{\hat{D}}(\boldsymbol{\theta}_{t-1}).$$

Let $\hat{\epsilon}(\boldsymbol{\theta}) = \nabla L(\boldsymbol{\theta}) - h_{\hat{D}}(\boldsymbol{\theta})$. By Lemma 7, there exists a sequence of $(N, N_s, N_d) \to \infty$ such that $\hat{\epsilon} = \sup_{\boldsymbol{\theta}} \|\hat{\epsilon}(\boldsymbol{\theta})\| \xrightarrow{p} 0$. Then we have

$$\begin{aligned}
-\eta h_{\hat{D}}(\boldsymbol{\theta}_{t-1})^\top \nabla L(\theta_{t-1}) &= -\eta h_{\hat{D}}(\boldsymbol{\theta}_{t-1})^\top \left(h_{\hat{D}}(\boldsymbol{\theta}_{t-1}) + \hat{\epsilon}(\boldsymbol{\theta}_{t-1})\right) \\
&\leq \eta \left(-\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 + (\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 + \|\hat{\epsilon}\|^2)/2\right) \\
&= -\frac{\eta}{2}\left(\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 - \hat{\epsilon}^2\right) \\
&\leq -\frac{\eta}{4}\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2,
\end{aligned}$$

under the case where $\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 \geq 2\hat{\epsilon}^2$. We note that

$$\begin{aligned}
L(\boldsymbol{\theta}_t) &\leq L(\boldsymbol{\theta}_{t-1}) - \frac{\eta}{4}\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 + \frac{\eta^2 \ell_0}{2}\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 \\
&\leq L(\boldsymbol{\theta}_{t-1}) - \frac{\eta}{8}\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2,
\end{aligned}$$

when $\eta < 1/4\ell_0$ which can be satisfied with a sufficiently small learning rate.

By summing over $t = 1, \ldots, T$, we have

$$L(\boldsymbol{\theta}_T) \leq L(\boldsymbol{\theta}_0) - 0.125\eta \sum_{t=1}^{T} \|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2.$$

Note that $L(\boldsymbol{\theta})$ is lower bounded by 0. Then we have $\sum_t \|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 = O(1)$. Thus there exists t in $\{0, \ldots, T\}$ such that $\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\|^2 = O(1/T)$.

Otherwise there exists t such that $\|h_{\hat{D}}(\boldsymbol{\theta}_{t-1})\| < \sqrt{2}\hat{\epsilon} = o_p(1)$.

Therefore we have the empirical estimator $\|h_{\hat{D}}(\hat{\boldsymbol{\theta}})\| \xrightarrow{p} 0$.

By the uniform convergence (10) from Lemma 7, we have $\|\nabla L(\hat{\boldsymbol{\theta}})\| = 0$. Then by the PL condition, there exists a sequence of $(N, N_s, N_d) \to \infty$ such that

$$L(\hat{\theta}) - L^* \xrightarrow{p} 0,$$

which leads to the desired result. ■

A.6 Proof of Lemma 6

We follow the same proof scheme as in Shen et al. (2020) where the only difference lies in the gradient with respect to the prior parameter β. To make this paper self-contained, we restate some proof steps here using our notations.

Let $\|\cdot\|$ denote the vector 2-norm. For a scalar function $h(x, y)$, let $\nabla_x h(x, y)$ denote its gradient with respect to x. For a vector function $g(x, y)$, let $\nabla_x g(x, y)$ denote its Jacobi matrix with respect to x. Given a differentiable vector function $g(x) : \mathbb{R}^k \to \mathbb{R}^k$, we use $\nabla \cdot g(x)$ to denote its divergence, defined as

$$\nabla \cdot g(x) := \sum_{j=1}^{k} \frac{\partial [g(x)]_j}{\partial [x]_j},$$

where $[x]_j$ denotes the j-th component of x. We know that

$$\int \nabla \cdot g(x)dx = 0$$

for all vector function $g(x)$ such that $g(\infty) = 0$. Given a matrix function $w(x) = (w_1(x), \ldots, w_l(x)) : \mathbb{R}^k \to \mathbb{R}^{k\times l}$ where each $w_i(x), i = 1 \ldots, l$ is a k-dimensional differentiable vector function, its divergence is defined as $\nabla \cdot w(x) = (\nabla \cdot w_1(x), \ldots, \nabla \cdot w_l(x))$.

To prove Lemma 6, we need the following lemma which specifies the dynamics of the generator joint distribution $p_g(x, z)$ and the encoder joint distribution $p_e(x, z)$, denoted by $p_\theta(x, z)$ and $p_\phi(x, z)$ here.

Lemma 13 *Using the definitions and notations in Lemma 6, we have*

$$\nabla_\theta p_{\theta,\beta}(x, z) = -\nabla_x p_{\theta,\beta}(x, z)^\top g_\theta(x) - p_{\theta,\beta}(x, z)\nabla \cdot g_\theta(x), \tag{19}$$

$$\nabla_\phi q_\phi(x, z) = -\nabla_z q_\phi(x, z)^\top e_\phi(z) - q_\phi(x, z)\nabla \cdot e_\phi(z), \tag{20}$$

$$\nabla_\beta p_{\theta,\beta}(x, z) = \nabla_x p_{\theta,\beta}(x, z)^\top \tilde{f}_\beta(x) - \nabla_z p_{\theta,\beta}(x, z)^\top f_\beta(z) - p_{\theta,\beta}(x, z)\nabla \cdot \begin{pmatrix}\tilde{f}_\beta(x)\\ f_\beta(z)\end{pmatrix}, \tag{21}$$

for all data x and latent variable z, where $g_\theta(G_\theta(z, \epsilon)) = \nabla_\theta G_\theta(z, \epsilon)$, $e_\phi(E_\phi(x, \epsilon)) = \nabla_\phi E_\phi(x, \epsilon)$, $f_\beta(F_\beta(\epsilon)) = \nabla_\beta F_\beta(\epsilon)$, and $\tilde{f}_\beta(G(F_\beta(\epsilon))) = \nabla_\beta G(F_\beta(\epsilon))$.

Proof [Proof of Lemma 13] We only prove (21) which is the distinct part from Shen et al. (2020).

Let l be the dimension of parameter β. To simplify notation, let random vector $Z = F_\beta(\epsilon)$ and $X = G(Z) \in \mathbb{R}^d$ and $Y = (X, Z) \in \mathbb{R}^{d+k}$, and let p be the probability density of Y. For each $i = 1, \ldots, l$, let $\Delta = \delta e_i$ where e_i is a l-dimensional unit vector whose i-th component is one and all the others are zero, and δ is a small scalar. Let $Z' = F_{\beta+\delta}(\epsilon)$, $X' = G(Z')$ and $Y' = (X', Z')$ so that Y' is a random variable transformed from Y by

$$Y' = Y + \begin{pmatrix}\tilde{f}_\beta(X)\\ f_\beta(Z)\end{pmatrix}\Delta + o(\delta).$$

Let p' be the probability density of Y'. For an arbitrary $y' = (x', z') \in \mathbb{R}^{d+k}$, let $y' = y + \binom{\tilde{f}_\beta(x)}{f_\beta(z)}\Delta + o(\delta)$ and $y = (x, z)$. Then we have

$$\begin{aligned}
p'(y') &= p(y)|\det(dy'/dy)|^{-1}\\
&= p(y)|\det(I_d + (\nabla\tilde{f}_\beta(x), \nabla f_\beta(z))^\top\Delta + o(\delta))|^{-1}\\
&= p(y)(1 + \Delta^\top\nabla \cdot (\tilde{f}_\beta(x), f_\beta(z))^\top + o(\delta))^{-1}\\
&= p(y)(1 - \Delta^\top\nabla \cdot (\tilde{f}_\beta(x), f_\beta(z))^\top + o(\delta))\\
&= p(y) - \Delta^\top p(y')\nabla \cdot (\tilde{f}_\beta(x'), f_\beta(z'))^\top + o(\delta)\\
&= p(y') - \Delta^\top(\tilde{f}_\beta(x'), f_\beta(z')) \cdot \nabla_{x'} p(x', z) - \Delta^\top p(y')(\nabla \cdot \tilde{f}_\beta(x'), \nabla \cdot f_\beta(z'))^\top + o(\delta).
\end{aligned}$$

Since y' is arbitrary, above implies that

$$\begin{aligned}
p'(x, z) = p(x, z) &- \Delta^\top(\tilde{f}_\beta(x), f_\beta(z)) \cdot (\nabla_x p(x, z), \nabla_z p(x, z))^\top \cdot \nabla_x p(x, z)\\
&- \Delta^\top p(x, z)(\nabla \cdot \tilde{f}_\beta(x'), \nabla \cdot f_\beta(z'))^\top + o(\delta)
\end{aligned}$$

for all $x \in \mathbb{R}^d, z \in \mathbb{R}^k$ and $i = 1, \ldots, l$, leading to (21) by taking $\delta \to 0$, and noting that $p = p_\beta$ and $p' = p_{\beta+\Delta}$. Similarly we can obtain (19) and (20). ■

Proof [Proof of Lemma 6] Recall the objective $D_{\mathrm{KL}}(q, p) = \int q(x, z) \log(p(x, z)/q(x, z))dxdz$. Denote its integrand by $\ell(q, p)$. Let $\ell_2'(q, p) = \partial \ell(q, p)/\partial p$. We have

$$\nabla_\beta \ell(q(x, z), p(x, z)) = \ell_2'(q(x, z), p(x, z))\nabla_\beta p_{\theta,\beta}(x, z)$$

where $\nabla_\beta p_{\theta,\beta}(x, z)$ is computed in Lemma 13.

Besides, we have

$$\begin{aligned}
\nabla_x \cdot [\ell_2'(q, p)p(x, z)\tilde{f}_\beta(x)] &= \ell_2'(q, p)p(x, z)\nabla \cdot \tilde{f}_\beta(x) \\
&\quad + \ell_2'(q, p)\nabla_x p(x, z) \cdot \tilde{f}_\beta(x) \\
&\quad + \nabla_x \ell_2'(q, p)p(x, z)\tilde{f}_\beta(x), \\
\nabla_z \cdot [\ell_2'(q, p)p(x, z)f_\beta(z)] &= \ell_2'(q, p)p(x, z)\nabla \cdot f_\beta(z) \\
&\quad + \ell_2'(q, p)\nabla p(x, z) \cdot f_\beta(z) \\
&\quad + \nabla \ell_2'(q, p)p(x, z)f_\beta(z).
\end{aligned}$$

Thus,

$$\nabla_\beta L_{\text{gen}} = \int \nabla_\beta \ell(q(x, z), p(x, z))dxdz = \int p(x, z)[\nabla_x \ell_2'(q, p)\tilde{f}_\beta(x) + \nabla_z \ell_2'(q, p)f_\beta(z)]$$

where we have $\nabla_x \ell_2'(q, p) = s(x, z)\nabla_x D^*(x, z)$ and $\nabla_x \ell_2'(q, p) = s(x, z)\nabla_z D^*(x, z)$.

Hence

$$\begin{aligned}
\nabla_\beta L_{\text{gen}} &= -\mathbb{E}_{(x,z)\sim p(x,z)} \left[s(x, z)(\nabla_x D^*(x, z)^\top \tilde{f}_\beta(x) + \nabla_z D^*(x, z)^\top f_\beta(z)) \right] \\
&= -\mathbb{E}_\epsilon \left[s(x, z)(\nabla_x D^*(x, z)^\top \nabla_\beta G(F_\beta(\epsilon)) + \nabla_z D^*(x, z)^\top \nabla_\beta F_\beta(\epsilon))|_{z=F_\beta(\epsilon)}^{x=G(F_\beta(\epsilon))} \right].
\end{aligned}$$

where the second equality follows from reparametrization. ■

Appendix B. Causal disentanglement and downstream tasks

In the main text, we first demonstrate the good performance of DEAR in causal disentanglement through causal controllable generation in Section 5.1, and then show the advantages of the DEAR representations in downstream tasks in terms of sample efficiency (Section 5.2.1) and distributional robustness (Section 5.2.2). In comparison with previous methods, majorly the VAE-based disentanglement methods, we adopt the same network architectures for the encoder and decoder, and use the same amount of annotated labels. In addition, for GraphVAE, we also assume the same prior information on the graph structure as DEAR. Therefore, we conclude that the superior performance of DEAR is due to better modeling. To further justify whether such advantages come from the disentanglement of the learned representations, in this section, we propose a metric for causal disentanglement based on the FactorVAE metric, and investigate the correlation between the disentanglement metric and the metrics for downstream tasks.

B.1 Metric for causal disentanglement

Many existing disentanglement papers also propose their metrics for disentanglement, including the β-VAE metric (Higgins et al., 2017), the FactorVAE metric (Kim and Mnih, 2018), the Mutual Information Gap (MIG) (Chen et al., 2018), the Separated Attribute Predictability (SAP) score (Kumar et al., 2018), etc. We refer the reader to Locatello et al. (2019) for a comprehensive introduction and discussion on these metrics.

However, all of these metrics only apply to the case where the ground-truth generative factors are mutually independent and do not apply when the factors are correlated. For example, the MIG score measures for each factor the normalized gap in mutual information with the highest and second highest coordinate in $E(x)$. Suppose a factor ξ_1 is correlated with ξ_2 and a disentangled representation $E(x)$ so that there exists 1-1 functions g_1 and g_2 such that $E_1(x) = g_1(\xi_1)$ and $E_2(x) = g_2(\xi_2)$. Then the mutual information of ξ_1 with (supposedly the highest coordinate) $E_1(x)$ and (supposedly the second highest coordinate) $E_2(x)$ will both be large, and then their difference will be small. As such, a disentangled representation in this case will not correspond to a large MIG score as expected.

To this end, we propose a metric for causal disentanglement (i.e., disentanglement of causally related ground-truth factors) based on the FactorVAE metric. Suppose there are m generative factors of interest $\xi_1, \dots, \xi_m$ which are causally related following the true SCM $\mathfrak{C}$ which is available. The procedure to compute the metric is presented in Algorithm 2. These steps largely follow those of the FactorVAE metric with the distinct parts tailored for causal disentanglement which we explain below the algorithm.

Algorithm 2: Metric for causal disentanglement

Input: Encoder E, meta-parameters M, N

1 **for** $k = 1, \dots, m$ **do**
2 **for** $i = 1, \dots, M$ **do**
3 Fix ξ_k to a randomly sampled value.
4 Randomly sample other factors ξ_{-k} from $\mathfrak{C}$ conditioning on ξ_k for N times.
5 Generate data with the N factors.
6 Obtain their representations using the learned encoder.
7 Normalize each dimension by its empirical standard deviation over the full data (or a large enough random subset).
8 Compute the empirical variance in each dimension of these normalized representations.
9 Take the index of the dimension with the lowest variance.
10 If the index matches k, it counts as a correct sample.
11 Let C_k be the total number of correct samples among the M samples.
12 Obtain score $S = \sum_{k=1}^{m} C_k/(KM)$.

Return: S

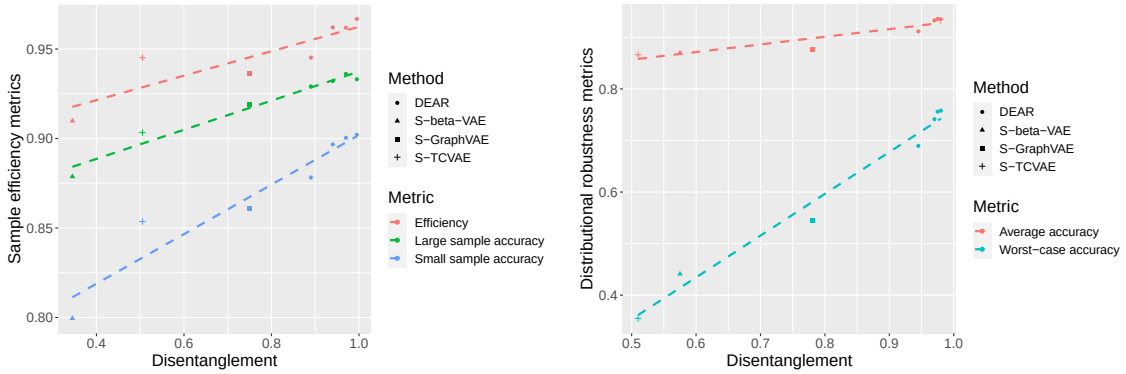
- Line 4: FactorVAE metric samples all factors independently from Uniform distributions, which does not match (and can be far away) from the true distribution of the causal factors. Instead, we sample the factors following the true SCM and hence respect the data distribution.

- Lines 10-12: FactorVAE metric uses the error rate of the majority-vote classifier as the metric, because in an unsupervised setting, one does not know which factor each representation captures. In contrast, the weakly-supervised setting can guarantee the alignment between each representation and a particular factor. Thus, we do not need the majority-vote classifier to identify this correspondence. Instead, we directly check whether the dimension with the lowest empirical variance matches the given index k.

As we notice, this metric is limited in that it not only requires the ground-truth factors of data for sufficient coverage of the data distribution as previous metrics do, but also requires the ground-truth SCM, which only happens in synthetic data. Nevertheless, in this work, we only use such a metric to provide evaluations and justification on the relationship between causal disentanglement and performance in downstream tasks. We leave a widely-applied quantitive metric for causal disentanglement to future work.

B.2 Experimental results

Figure 14 shows the scatter plots of the metrics that we considered in downstream tasks (Section 5.2) and the metric for causal disentanglement (with $M = 200$ and $N = 50$). Each metric is used to evaluate seven disentanglement models, including S-β-VAE, S-TCVAE, S-GraphVAE, and multiple DEAR-LIN models with $\lambda = 0.1, 1, 5, 10$. All models are trained using fully supervised labels and GraphVAE and DEAR are given the true graph structure. The network architectures for the encoders and decoders are all the same. We observe a positive correlation between causal disentanglement and performance in downstream tasks, which indicates that the learned representations with a higher disentanglement score tend to perform better in terms of sample efficiency and distributional robustness in downstream tasks. In particular, we notice that the small sample accuracy and worst-case accuracy benefit the most from better causal disentanglement for the corresponding fitted lines have the largest scope.



(a) Sample efficiency

(b) Distributional robustness

Figure 14: Relationship between causal disentanglement and performance in downstream tasks.

Appendix C. Discussion: supervision for disentanglement learning

We comment on the two forms of supervision that may be available and commonly considered in literature for the task of disentangled representation learning.

- Form 1 (direct and few labels): in some scenarios, we may have some conceptual knowledge about the data in the sense that we know the concepts of the underlying generative factors of data, especially those concepts that we are interested in. In such cases, a weakly supervised setting is feasible where only a few samples have annotated labels of the factors, since at least manual labeling of a few examples is practical. A representative work uses this form of supervision is Locatello et al. (2020b).

- Form 2 (auxiliary information of a full sample): in some other scenarios, we have no prior knowledge on what the ground-truth concepts are and thus cannot get the direct annotated labels of them. Auxiliary information is then needed for all samples with some assumptions on the variability of such side information as well as its correlation with the true generative factors. A representative work along this line is Khemakhem et al. (2020).

Both settings have some real applications and limitations which make them complementary. On one hand, Form 2 in general tends to require "weaker" supervision than Form 1 in the sense that it does not require direct annotations of the true factors themselves. Thus, efforts towards general provable disentanglement should be put in studying along Form 2. However, in fact, the auxiliary observed variables in Form 2 also require certain knowledge on the true factors in order to verify the mathematical assumptions required in identifiability, e.g. the variability condition in Khemakhem et al. (2020). Intuitively, the auxiliary variables which can guarantee disentanglement should have enough variability and correlation with the true factors. In addition, current identifiability theory with Form 2 still assumes relatively strong and limited structure assumption on the true factors, e.g., conditional independence in Khemakhem et al. (2020).

On the other hand, current research on disentanglement mostly focuses on the scenarios where we indeed have some conceptual knowledge on the true factors, which makes Form 1 at least a feasible and practical setting. For simple structures of true factors (e.g., independence or conditional independence, as assumed in most previous work), existing methods with Form 1 can achieve disentanglement, which is much more straightforward compared to provable disentanglement with supervision of Form 2. However, for more complex structures (e.g., a causal graph, as considered in our paper), existing methods using independent or conditionally independent priors generally cannot identify disentanglement even with supervision in Form 1, as shown in our Proposition 4. In particular, existing formulations (e.g., Locatello et al. (2020b)) in general cannot even reach the optimum of the supervised loss, so they cannot disentangle. To this end, our paper proposes a bidirectional generative model with an SCM prior trained using a GAN-type algorithm, which resolves this problem under the clearly stated setup and assumptions.

Appendix D. Discussion: generalization to unseen interventions

We recall that DEAR is trained on observational data, that is, the training data is IID sampled from the data distribution q_x and the latent variables follow IID a joint distribution p_z, e.g., induced by a SCM, without a mixture with interventional distributions. When the generative model is perfectly learned, we have $q_x(x) = \int p_{G^*}(x|z)p_z(z)dz$. Then an interesting question would be how our method generalizes to unseen interventions. Specifically let $p_z^I(z)$ be an interventional distribution. The consequent data distribution $q_x^I(x) = \int p_{G^*}(x|z)p_z^I(z)dz$ does not match the observational distribution q_x and model trained on an IID sample from q_x have not seen q_x^I.

Now we give some insights on how given the true graph structure, DEAR trained on observational data can sample from an interventional distributions q_x^I. We start with the general definition of SCM. A structural causal model (SCM) over variables $Z_i, i = 1 \ldots, m$ can be generally expressed as

$$Z_i = f_i(Pa(Z_i; A), \epsilon_i),\ i = 1, \ldots, m, \tag{22}$$

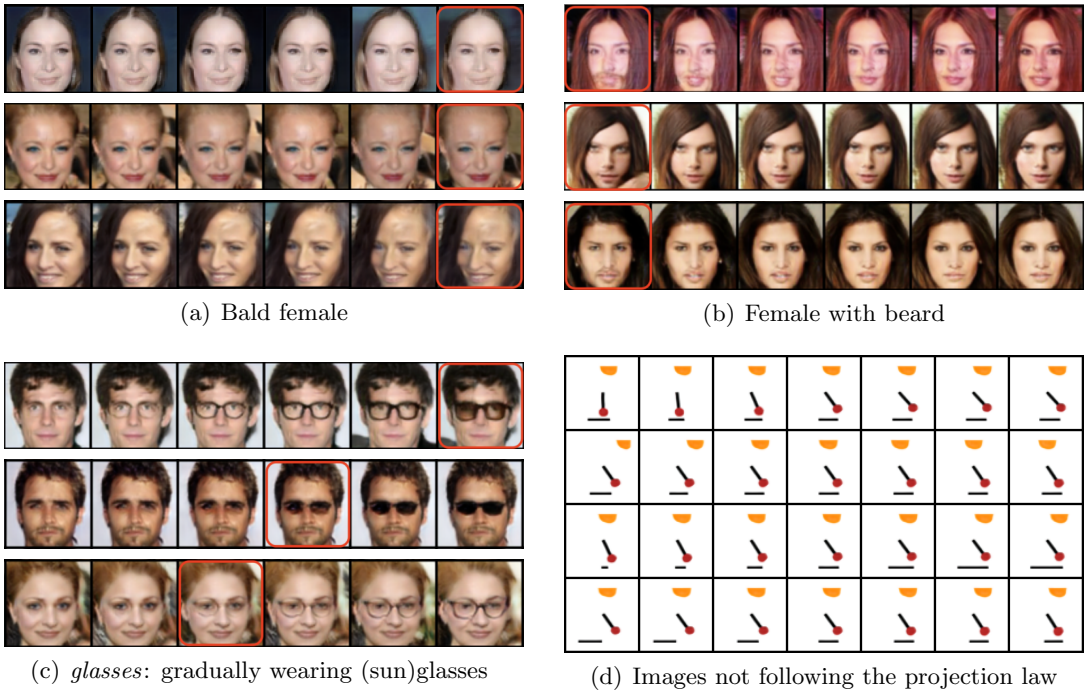
where A denotes the adjacency matrix, $Pa(Z_i; A)$ denotes the set of parents of node Z_i, and ϵ_i is the exogenous noise. Learning of an SCM consists of structure learning of A and parameter estimation of all the assignments $f_i, i = 1, \ldots, m$, in the SCM, i.e., how each node is generated given its parents and exogenous noise. When given the underlying causal structure, standard parameter estimation methods like maximum likelihood estimation can yield a consistent estimator of the true SCM assignments from the observational data:

$$Z_i = \hat{f}_i(Pa(Z_i; A), \epsilon_i),\ i = 1, \ldots, m. \tag{23}$$

Note that an intervention can be defined as operations that modify a subset of assignments in (22), e.g., changing ϵ_i, or setting f_i (and thus Z_i) to a constant (Pearl et al., 2000; Schölkopf, 2019). Therefore, with the estimated SCM (23) at hand, we can sample from any interventional distributions.

We illustrate this through some experimental results shown in Figure 15. In (a), we intervene on the two factors *bald* and *gender*. In each line, we keep *gender* = female and gradually increase the probability of them being bald. Particularly in the red box, we obtain images of bald female faces which have never been seen from the observational data. In (b), we intervene on *beard* and *gender* to generate images of female with beard which are shown in the red box. In (c), we show some generated samples that gradually wear (sun)glasses, while in the training data, there are only images with or without glasses but no intermediate states. In (d), we intervene on all four factors. In each line, the image in the middle follows the true SCM (described later in Appendix F) so that the factors satisfy the projection law. Then we change the value of only one factor while keeping others fixed, which leads to samples not satisfying the projection law. In summary, we see that although these interventions are not appearing in the observational data, DEAR is able to generate samples from such interventional distributions, suggesting its generalizability to unseen interventions.

More systematic analysis on the out-of-distribution generalizability of the encoder is to be explored in future work. One potential direction is to utilize the generalizability of the generator to unseen interventions to improve the OOD performance of the encoder. Along



(a) Bald female

(b) Female with beard

(c) *glasses*: gradually wearing (sun)glasses

(d) Images not following the projection law

Figure 15: Samples from unseen interventional distributions.

this direction, for example, Sauer and Geiger (2021) recently combined disentangled generative models and out-of-distribution classification, but adopted a different disentanglement framework.

Appendix E. Experiments in the independent case

In this section, we test our method on benchmark data sets where the ground-truth generative factors are independent, which is a spacial case of the causal case with no edge in the graph structure. Gondal et al. (2019) proposed a real-world benchmark data set, MPI3D-real, which consists of over one million images of physical 3D objects with seven independent factors of variation such as object color, shape, size and position. They also provided two simulation data sets. We test a simplified version of DEAR with an independent prior (a standard Gaussian) on real data MPI3D-real and simulated data MPI3D-simu (MPI3D-realistic in Gondal et al. (2019)). Both data sets consist of 1,036,800 images with resolution $64 \times 64 \times 3$. We assume 0.01% of the data (around 100 samples) have annotated labels. No prior information on the graph structure is needed since we directly use an independent prior for the latent variable instead of an SCM.

We are interested in the disentanglement performance on both real and simulated data, as well as the transferability of the representations from simulation to the real-world or reverse. As shown in the experiments by Gondal et al. (2019), most existing VAE-based methods perform similarly in disentanglement and all the metric for disentanglement also gives similar results. Hence, we consider weakly-supervised TCVAE as a representative of

the baseline methods and consider FactorVAE metric to measure disentanglement. As we have mentioned, the weakly-supervised setting can guarantee the alignment between each representation and a particular factor. Therefore, when computing the FactorVAE metric, we skip the majority-vote classifier and directly apply lines 10-12 in Algorithm 2 to obtain the score.

As shown in Table 3, DEAR always significantly outperforms TCVAE in the disentanglement score and is particularly superior when training and testing on the same data set. In the transfer setting where we apply the encoder trained on one data set to another data set, both methods suffer from a performance decline. This is consistent with the discovery in Gondal et al. (2019) who found that direct transfer of learned representations from simulated to real data seems to work rather poorly. To sum up, this section suggests that DEAR can achieve state-of-the-art performance in data whose underlying factors are independent, though it is developed to handle the causal case.

Train	Test	Method	Disentangle
MPI3D-simu	MPI3D-simu	DEAR	0.9543
		TCVAE	0.5800
MPI3D-real	MPI3D-real	DEAR	0.9579
		TCVAE	0.5793
MPI3D-simu	MPI3D-real	DEAR	0.4879
		TCVAE	0.3614
MPI3D-real	MPI3D-simu	DEAR	0.5571
		TCVAE	0.3443

Table 3: Results on MPI3D data.

Appendix F. Implementation details

In this section, we provide the details of the experimental setup and the network architectures used for all experiments, followed by a description of the synthesized Pendulum data set.

Preprocessing and hyperparameters. We pre-process the images by taking center crops of 128×128 for CelebA and resizing all images in CelebA and Pendulum to the 64×64 resolution. We adopt Adam with $\beta_1 = 0$, $\beta_2 = 0.999$, and a learning rate of 1×10^{-4} for D, 5×10^{-5} for E, G and F, and 1×10^{-3} for the weighted adjacency matrix A. We use a mini-batch size of 128. For adversarial training in Algorithm 1, we train the D once on each mini-batch. The coefficient λ of the supervised regularizer is set to 5 unless indicated otherwise. We use CE supervised loss for both CelebA with binary observations of the underlying factors and Pendulum with bounded continuous observations. Note that L_2 loss works comparable to CE loss on Pendulum. The results of DEAR and baseline methods in controllable generation presented in Section 5.1 and Appendix G use full supervision of underlying generative factors, i.e., $N_s = N$, since the qualitative results with 10% labels have no big difference.

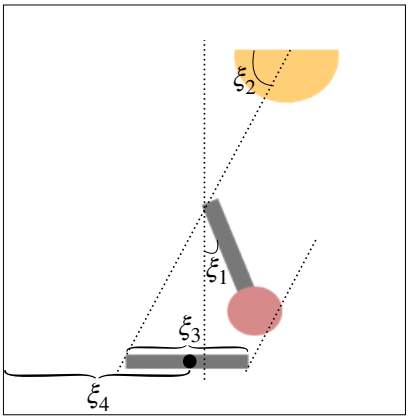


Figure 16: Generative factors of the Pendulum data set. ξ_1: *pendulum_angle*, ξ_2: *light_angle*, ξ_3: *shadow_length*, ξ_4: *shadow_position*.

In downstream tasks, for BGMs with an encoder, we train a two-level MLP classifier with 100 hidden nodes using Adam with a learning rate of 1×10^{-2} and a mini-batch size of 128. Models were trained for around 150 epochs on CelebA, 600 epochs on Pendulum, and 50 epochs on MPI3D on NVIDIA RTX 2080 Ti.

Description of the Pendulum data set. In Figure 16, we illustrate the generative factors of the synthesized Pendulum data set, following Yang et al. (2021). Given the *pendulum_angle*(ξ_1) and *light_angle*(ξ_2), following the projection law, one can determine the *shadow_length*(ξ_3) and *shadow_position*(ξ_4). Note that we consider the parallel light in our simulator. Specifically, define some constants: $c_x = 10, c_y = 10.5$ are the axis's of the center (pendulum origin); $l_p = 9.5$ be the pendulum length (including the red ball); the bottom line of a single plot corresponds to $y = b$ with base $b = -0.5$. Then the ground-truth structural causal model is expressed as follows.

$$
\begin{aligned}
\xi_1 &\sim \mathcal{U}(\pi/4, \pi/2) \\
\xi_2 &\sim \mathcal{U}(0, \pi/4) \\
\xi_3 &= \left(c_x + l_p \sin \xi_1 - \tfrac{c_y - l_p \cos \xi_1 - b}{\tan \xi_2}\right) - \left(c_x - \tfrac{c_y - b}{\tan \xi_2}\right) \\
\xi_4 &= \left(c_x + l_p \sin \xi_1 - \tfrac{c_y - l_p \cos \xi_1 - b}{\tan \xi_2} + c_x - \tfrac{c_y - b}{\tan \xi_2}\right) / 2.
\end{aligned}
$$

where $\mathcal{U}(a, b)$ denotes the uniform distribution on interval (a, b).

Implementation of the SCM. Recall the nonlinear SCM as the prior

$$z = f((I - A^\top)^{-1} h(\epsilon)) := F_\beta(\epsilon).$$

We find Gaussians are expressive enough as unexplained noises, so we set h as the identity mapping. As mentioned in Section 4.1 we require the invertibility of f. We implement both linear and nonlinear ones. For a linear f, we formally refer to $f(z) = Wz + b$, where W and

b are learnable weights and biases. Note that W is a diagonal matrix to model the element-wise transformation. Its inverse function can be easily computed by $f^{-1}(z) = W^{-1}(z - b)$.

For a non-linear f, we use piece-wise linear functions defined by

$$[f([z]_i)]_i = [w_0]_i[z]_i + \sum_{t=1}^{N_a} [w_t]_i([z]_i - a_i)\mathbf{I}([z]_i \geq a_i) + [b]_i$$

where $a_0 < a_1 < \cdots < a_{N_a}$ are the points of division, $\mathbf{I}(\cdot)$ is the indicator function, and $\{b, w_t : t = 0, \ldots, N_a\}$ is the set of learnable parameters. According to the denseness of piecewise linear functions in $C[0, 1]$ (Shekhtman, 1982), the family of such piece-wise linear functions is expressive enough to model general element-wise non-linear invertible transformations.

Network architectures. We follow the architectures used in Shen et al. (2020). Specifically, for such realistic data, we adopt the SAGAN (Zhang et al., 2019) architecture for D and G. The D network consists of three modules as shown in Figure 17(a) and detailed described in Shen et al. (2020). Architectures for network G and D_x are given in Figure 17(b-c) and Table 4. The encoder architecture is the ResNet50 (He et al., 2016) followed by a 4-layer MLP of size 1024 after ResNet's global average pooling layer.

Table 4: SAGAN architecture ($k = 100$ for CelebA and $k = 6$ for Pendulum and $ch = 32$).

(a) Generator

Input: $z \in \mathbb{R}^k \sim p_z$
Linear $\to 4 \times 4 \times 16ch$
ResBlock up $16ch \to 16ch$
ResBlock up $16ch \to 8ch$
ResBlock up $8ch \to 4ch$
Non-Local Block (64×64)
ResBlock up $4ch \to 2ch$
BN, ReLU, 3×3 Conv $2ch \to 3$
Tanh

(b) Discriminator module D_x

Input: RGB image $x \in \mathbb{R}^{64 \times 64 \times 3}$
ResBlock down $ch \to 2ch$
Non-Local Block (64×64)
ResBlock down $2ch \to 4ch$
ResBlock down $4ch \to 8ch$
ResBlock down $8ch \to 16ch$
ResBlock $16ch \to 16ch$
ReLU, Global average pooling (f_x)
Linear $\to 1$ (s_x)

Experimental details for baseline methods. We reproduce the S-VAEs including S-VAE, S-β-VAE and S-TCVAE using E and G with the same architectures as DEAR's and adopt the same optimization algorithm with same hyperparameters for training. The coefficient for the independence regularizer is set to 4 since we notice that setting a larger independence regularizer hurts disentanglement in the correlated case. We implement GraphVAE by ourselves using the same architectures (for the encoder and decoder) and optimizer as DEAR. The latent dependencies of GraphVAE consists of a bottom-up network (approximate $z|x$):

nn.Linear(latent_dim, 32), nn.BatchNorm1d(32), nn.ELU(),
nn.Linear(32, node_dim), nn.Linear(node_dim, 2*node_dim)

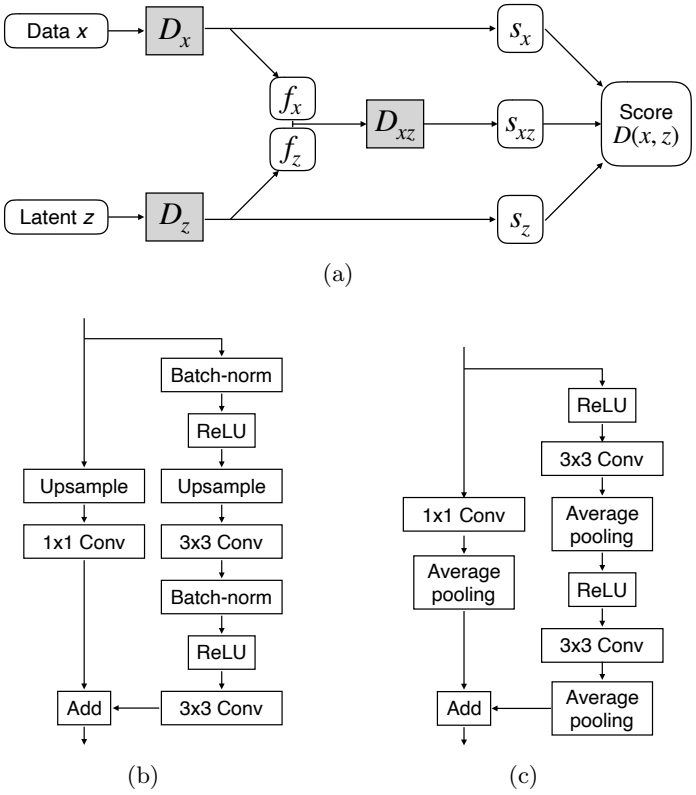


Figure 17: (a) Architecture of the discriminator $D(x, z)$; (b) A residual block (up scale) in the SAGAN generator where we use nearest neighbor interpolation for *Upsampling*; (c) A residual block (down scale) in the SAGAN discriminator.

and a top-down network (approximate $z|$parent):

nn.Linear(n_parent_nodes *node_dim, 32), nn.BatchNorm1d(32), nn.ELU(),
nn.Linear(32, node_dim), nn.Linear(node_dim, 2*node_dim).

Note that this implementation follows the original one: $z|x,$ parent is obtained by precision-weighted fusion in He et al. (2018). Since our factor dependency are explicit, we use 32 latent dimension for more efficient optimization.

For the supervised regularizer, we use $\lambda = 1000$ for a balance of generative modeling and supervised regularizer. The ERM ResNet is trained using the same optimizer with a learning rate of 1×10^{-4}. We run the public source code from `https://github.com/mkocaoglu/CausalGAN` to produce the results of CausalGAN.

Appendix G. Additional results in causal controllable generation

In this section, we present more qualitative results in causal controllable generation on two data sets using DEAR and baseline methods, including S-VAEs (Locatello et al., 2020b), GraphVAE (He et al., 2018), and CausalGAN (Kocaoglu et al., 2018). We consider three

underlying structures on two data sets: Pendulum in Figure 2(a), CelebA-Smile in Figure 2(b), and CelebA-Attractive in Figure 2(c). Note that the ordering of the rows in the traversals below matches the indices in Figure 2.

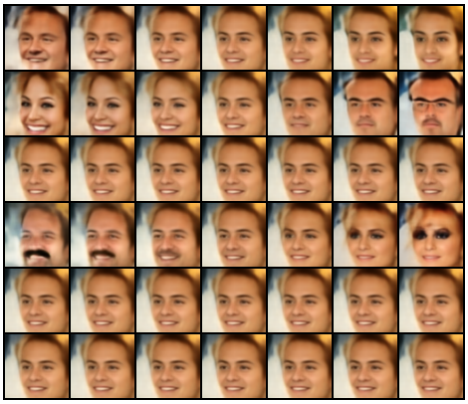
(a) Traversal (CelebA-Smile) (b) Intervention (CelebA-Smile)

(c) Traversal (CelebA-Attractive) (d) Intervention (CelebA-Attractive)

(e) Traversal (Pendulum) (f) Intervention (Pendulum)

Figure 18: Results of DEAR. On the left we present the traditional latent traversals (the first type of intervention stated in Section 5.1) which show the disentanglement. On the right we show the results of intervening on one latent variable from which we see the consequent changes of the others (the second type of intervention). Specifically intervening on the cause variable influences the effect variables while intervening on effect variables makes no difference to the causes.

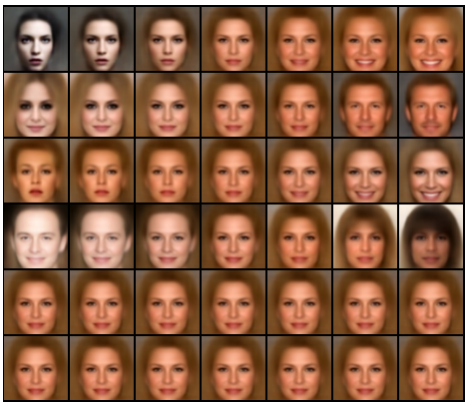
(a) S-TCVAE (CelebA-Smile)



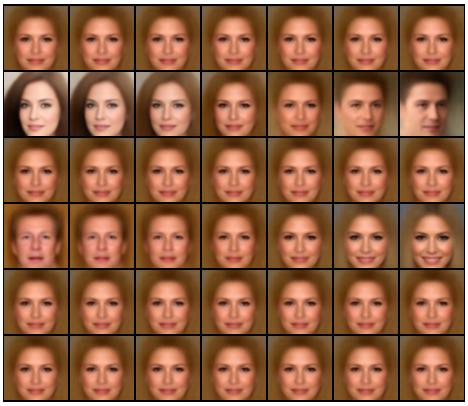
(b) S-TCVAE (CelebA-Attractive)

(c) S-FactorVAE (CelebA-Smile)

(d) S-FactorVAE (CelebA-Attractive)

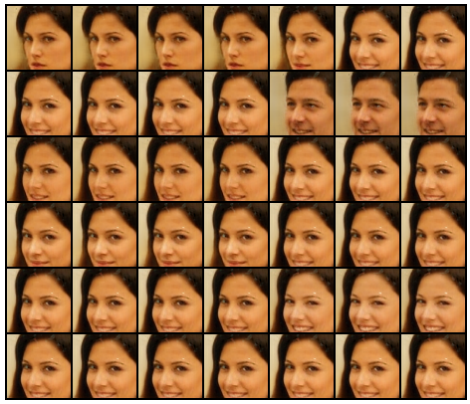


(e) S-β-VAE (CelebA-Smile)



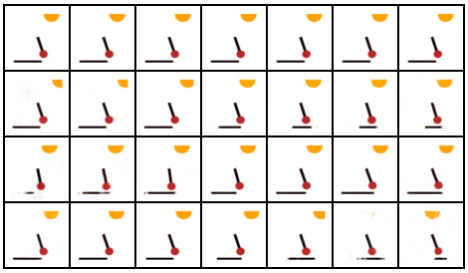
(f) S-β-VAE (CelebA-Attractive)

Figure 19: Traversal results of baseline methods. We see that entanglement occurs and some factors are not captured by the generative models (traversing on some dimensions of the latent vector makes no difference in the decoded images.) Besides, the generated images from VAEs are blurry.

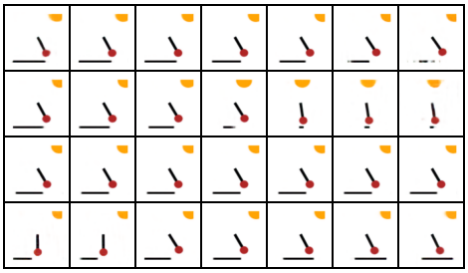


(a) CausalGAN (CelebA-Smile)

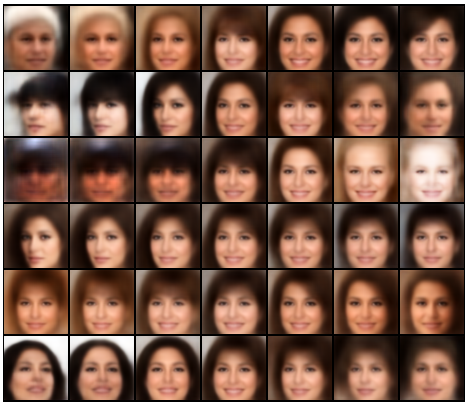
(b) CausalGAN (CelebA-Attractive)



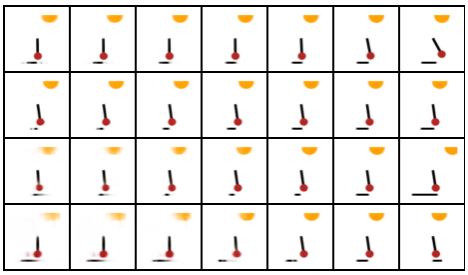
(c) S-TCVAE (Pendulum)



(d) S-FactorVAE (Pendulum)



(e) S-GraphVAE (CelebA-Attractive)



(f) S-GraphVAE (Pendulum)

Figure 20: Traversal results of baseline methods. CausalGAN uses the binary factors as the conditional attributes, so the traversals (a-b) appear some sudden changes. In contrast, we regard the continuous logit of binary labels as the underlying factors and hence enjoy smooth manipulations. In addition, the controllability of CausalGAN is also limited, since entanglement still exists. Results of S-VAEs are explained in Figure 19. The traversal of S-GraphVAE on Pendulum looks better than those of S-VAEs, especially in the first two factors, while the performance on CelebA is poor. Besides, S-GraphVAE has poor generation quality.

References

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

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.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxWIgBFPS`.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *arXiv preprint arXiv:2007.01754*, 2020.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nicholas Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *NeurIPS Workshop of Learning Disentangled Features*, 2017.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wuthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=8VXvj1QNRl1`.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2017.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martín Arjovsky, Olivier Mastropietro, and Aaron C. Courville. Adversarially learned inference. In *International Conference on Learning Representations*, 2017.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Muhammad Waleed Gondal, Manuel Wüthrich, DJordje Miladinović, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *arXiv preprint arXiv:1906.03292*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Jiawei He, Yu Gong, Joseph Marino, Greg Mori, and Andreas Lehrmann. Variational autoencoders with jointly optimized latent dependency structure. In *International Conference on Learning Representations*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Irina Higgins, Loïc 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. In *International Conference on Learning Representations*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Nan Rosemary Ke, Aniket Rajiv Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Jimenez Rezende, Michael Curtis Mozer, Yoshua Bengio, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. 2021.

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

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Felix Leeb, Yashas Annadani, Stefan Bauer, and Bernhard Schölkopf. Structural autoencoders improve representations for generation and transfer. *arXiv preprint arXiv:2006.07796*, 2020.

Zinan Lin, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *International Conference on Machine Learning*, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *arXiv preprint arXiv:2003.00307*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, June 2019. URL `http://proceedings.mlr.press/v97/locatello19a.html`.

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

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variation using few labels. In *International Conference on Learning Representations*, 2020b.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400. JMLR. org, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Raha Moraffah, Bahman Moraffah, Mansooreh Karami, Adrienne Raglin, and Huan Liu. Can: A causal adversarial network for learning observational and interventional distributions. *arXiv preprint arXiv:2008.11376*, 2020.

Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *arXiv preprint arXiv:2006.10201*, 2020.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 2000.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14866–14876, 2019.

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

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=BXewfAYMmJw`.

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

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *International Conference on Machine Learning*, 2012.

Boris Shekhtman. Why piecewise linear functions are dense in c [0, 1]. *Journal of Approximation Theory*, 36(3):265–267, 1982.

Xinwei Shen, Tong Zhang, and Kani Chen. Bidirectional generative modeling using adversarial gradient estimation. *arXiv preprint arXiv:2002.09161*, 2020.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 3738–3746, 2016.

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

Jan Stühmer, Richard Turner, and Sebastian Nowozin. Independent subspace analysis for unsupervised learning of disentangled representations. In *International Conference on Artificial Intelligence and Statistics*, pages 1200–1210. PMLR, 2020.

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*, pages 6056–6065. PMLR, 2019.

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*, pages 10401–10412. PMLR, 2021.

Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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*, pages 9593–9602, June 2021.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, 2019.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

Jiji Zhang and Peter Spirtes. Intervention, determinism, and the causal minimality condition. *Synthese*, 182(3):335–347, 2011.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from generative models. In *International Conference on Machine Learning*, 2017.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.