C-Disentanglement: Discovering Causally-Independent Generative Factors under an Inductive Bias of Confounder

Anonymous Authors[1]

Abstract

Representation learning assumes that real-world data is generated by a few causally disentangled generative factors (i.e., sources of variation). However, most existing works assume unconfoundedness (i.e., there are no common causes to the generative factors) in the discovery process, and thus obtain only statistical independence. In this paper, we recognize the importance of modeling confounders in discovering causal generative factors. Unfortunately, such factors are not identifiable without proper inductive bias. We fill the gap by introducing a framework named **C**onfounded-**Disentanglement** (C-Disentanglement), the first framework that explicitly introduces the inductive bias of confounder via labels/knowledge from domain expertise. We further propose an approach for sufficient identification under the VAE framework.

1. Introduction

Causally disentangled representation learning methods endeavor to identify and manipulate the underlying explanatory causes of variation (i.e., generative factors) within observational data through obtaining *causally disentangled representations* (Wang & Jordan, 2021; Eastwood & Williams, 2018). Pursuing causal independence[1] makes it possible to identify the ground truth generative factors that are not statistically independent, for example, the color, shape and size in a fruit dataset as shown in **??**, which is more realistic and allows more controlled data generation, improved robustness, and better generalization in out-of-distribution problems.

Despite great success, existing methods suffer from the identifiability issue in discovering semantically meaningful generative factors. The main reason is that most of them equate disentangling in the latent space with enforcing statistical independence (Higgins et al., 2016; Kim & Mnih, 2018; Chen et al., 2016) in the latent space, or requiring no mutual information(Eastwood & Williams, 2018). In other words, they explicitly or implicitly assume that the observational dataset is *unconfounded*, i.e., there are no common causes among the learned latent factors. Such a limited assumption leads to discrepancies in characterizing the statistical relationship of the generative factors on the observational set and makes them *non-identifiable*. In addition, it has been shown *almost impossible* to obtain disentangled representations through purely unsupervised learning without proper inductive biases (Horan et al., 2021; Locatello et al., 2019).

In this paper, we recognize the importance of providing inductive bias to confounder so that the ground truth generative factors can be identified and we fill the gap by introducing the inductive bias via knowledge from domain expertise. Specifically, we propose a framework called **C**onfounded-**Disentanglement** (*C-Disentanglement*). C-Disentanglement is, to the best of our knowledge, the first framework that discusses the identifiability issue of generative factors regarding the inductive bias of confounder, and thus opens up the possibility to discover the ground truth causally disentangled generative factors which are correlated in the observational dataset. Under the framework, We develop an algorithm to discover the causally disentangled generative factors in the latent space with inductive bias $\mathbf{C}$, where $\mathbf{C}$ is a label set.

Summary of contributions: (1) We recognize the identifiability issue of discovering generative factors in the latent space. We accordingly introduce a framework, named Confounded Disentanglement (C-Disentanglement). It is the first framework that discusses how inductive bias of confounder could be explicitly provided via labels/knowledge from domain expertise and propose an algorithm, cdVAE, to identify these factors in the latent space.

(2) We conduct extensive experiments and ablation studies across various datasets and tasks. Empirical results verify that cdVAE outperforms existing methods in terms of inferring causally disentangled latent representation and also

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SPIGM workshop at ICML 2023. Do not distribute.

[1]We will use causally disentangled and causally independent interchangeably.

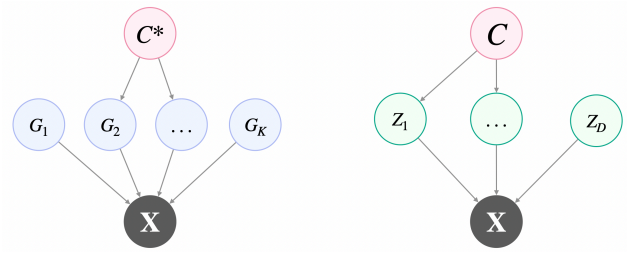

Figure 1: The left figure shows the ground truth generative process while the right figure demonstrates the learning task.

show cdVAE's superiority in downstream tasks under OOD generalization.

2. Confounded causal disentanglement via inductive bias

2.1. Problem formulation

We formally frame the task of discovering a set of generative factors from the observational dataset from a causal perspective as shown in Figure 1. Let $\mathbf{X}$ denote the observational data, the confounded causal generative process (Suter et al., 2019; Reddy et al., 2022) assumes that $\mathbf{X}$ are generated from K ground-truth causes of variations $G = [G_1, G_2, ..., G_K]$ (i.e., $G \to X$) that do not cause each other. These generative factors are generally not available, and they are confounded by some unobserved confounding variables $\mathbf{C}^*$. The task of interest is to discover those generative factors in the latent space (denoted as $\mathbf{Z} = [Z_1, Z_2, ..., Z_D]$) that best approximates $\mathbf{G}$ from $\mathbf{X}$.

Causal disentanglement among latent generative factors. Generative factors, encoded in the latent representational space, are causally disentangled, meaning that intervening on the value of one factor does not affect the distribution of the others. It is formally defined as:

Definition 2.1 (Causal Disentanglement on Data $\mathbf{X}$((Pearl, 2009; Suter et al., 2019; Wang & Jordan, 2021))). A representation is disentangled if, for $i \in \{1, ..., D\}$,

$$P(Z_i|\text{do}(Z_{-i} = z_{-i}), \mathbf{X}) = P(Z_i|\mathbf{X}), \quad \forall z_i. \tag{1}$$

where $-i = \{1, 2, ..., D\}/i$ indicates the set of all indices except for i.

Challenge of unobserved $\mathbf{C}^*$. The generative factors $\mathbf{G}$ are not identifiable without proper inductive bias of $\mathbf{C}^*$ (Locatello et al., 2019).

Previous works on discovering the generative factors either obtain disentanglement by enforcing statistical independence on latent variables (Higgins et al., 2016; Liu et al., 2015; Chen et al., 2016), or require that latent variables do not capture information of each other (Eastwood & Williams, 2018). Such a setting is equivalent to assuming unconfoundedness of the generative factors. It ignores the possibility that correlated latent variables can also be causally disentangled in the observational distribution, and hence is an assumption too restrictive. Fortunately, even though the ground truth generative factors are unobserved, domain expertise may inform a "reasonable" or "likely" inductive bias of the confounder from an accessible label set, denoted as $\mathbf{C}$.

We thus introduce an operator $do^c(\cdot)$ to estimate the interventional distribution on the observational set under inductive $\mathbf{C}$. This $\mathbf{C}$ is used to account for all correlation among $\mathbf{Z}$. The framework that applies $do^c(\cdot)$ is named C-Disentanglement.

We empirically show in Section 4 that even partial information of confounders improves accuracy, outperforming existing methods in various tasks, even under distribution shifts.

Definition 2.2 (C-Disentanglement and do^c)**.** Let $\mathbf{X}$ be the observational data, $\mathbf{Z} = [Z_1, Z_2, ..., Z_D]$ be a concatenation of D random variables, $\mathbf{C}$ be a label set selected from domain expertise to provide inductive bias for confounders of the observational data, we define do^c operation as:

$$P(Z_i|do^c(Z_{-i} = z_{-i}), \mathbf{X}) = \sum_{c \in \mathbf{C}} P(Z_i|\mathbf{X}, Z_{-i} = z_{-i}, \mathbf{C} = c)P(\mathbf{C} = c)$$

$\forall i \in 1, 2, 3, ..., D$, where c is realizations of $\mathbf{C}$. $\mathbf{Z}$ obtains C-Disentanglement on $\mathbf{X}$ given $\mathbf{C}$ if

$$P(Z_i|do^c(Z_{-i} = z_{-i}), \mathbf{X}) = P(Z_i|\mathbf{X}). \tag{2}$$

3. cdVAE: identify causally disentangled factors in the latent space.

In this section, we provide an algorithm, cdVAE, for confounded disentangled VAE, to identify the latent causally disentangled generative factors under confounder $\mathbf{C}$ in the context of VAE.

3.1. Learning objective

Given $\mathbf{X}$ as the dataset, we hope to find a deterministic function f, parameterized by θ, where $f : \mathcal{Z} \to \mathcal{X}$ such that (1) $P(X) = \int f(\mathbf{Z}; \theta)P(\mathbf{Z})dz$ is maximized and (2) each Z_i encodes causally disentangled generative factors, with $\mathbf{Z} = [Z_1, Z_2, ..., Z_D]$ as a concatenation of random variables in the latent space.

More concretely, $f(z; \theta)$ is characterized by a probability distribution $P(\mathbf{Z}|\mathbf{X}; \theta)$, and $p(\mathbf{Z})$ is a prior distribution from where $\mathbf{Z}$ can be easily sampled. For each $Z_i \in \mathbf{Z}$, we require

$$P(Z_i|do^c(Z_{-i}), \mathbf{X}) = P(Z_i|\mathbf{X}), \quad \forall c \in \mathbf{C}. \tag{3}$$

Applying the variational Bayesian method (Kingma & Welling, 2013), even though the methodology is not restricted to only VAEs, the learning object is to optimize the evidence lower bound (ELBO) while satisfying the constrain on causal disentanglement:

$$\max_{\theta,\phi} \quad \mathbb{E}_{\mathbf{Z}\sim Q}[logP(\mathbf{X}|\mathbf{Z};\theta)] - D[Q(\mathbf{Z}|\mathbf{X};\phi)||P(\mathbf{Z})] \tag{4}$$

$$s.t. \quad P(Z_i|do^c(Z_{-i}),\mathbf{X}) - P(Z_i|\mathbf{X}) = 0 \quad \forall i \in 1,2,...,D \tag{5}$$

For simplicity, we omit all model parameters θ, ϕ in writing.

3.2. Learning strategy

We start from the estimation of the causal disentanglement constraint as shown in Equation (3). Because the probability distribution is hard to be directly calculated, inspired by (Suter et al., 2019), we resort to its first-order moment as an approximation. Specifically, we estimate the L_1 distance between the expectation of probability $P(Z_i|do^c(Z_{-i}), \mathbf{X})$ and $P(Z_i|\mathbf{X})$ as follows:

$$l_c = \sum_{i=1}^{D}[\mathbb{E}(Z_i|do^c(Z_{-i}),\mathbf{X}) - \mathbb{E}(Z_i|\mathbf{X})]. \tag{6}$$

We show in the following theorem that restraining the learned latent variable $\mathbf{Z}$ to be a Gaussian distribution with diagonal covariance matrix minimizes loss l_c. Proof can be found in Appendix D.2.

Theorem 3.1. *Suppose that the latent variable* $\mathbf{Z}$ *on dataset* $\mathbf{X}$ *given* $\mathbf{C} = c$ *is Gaussian* $\mathcal{N}(\mu^c(\mathbf{X}), \Sigma^c(\mathbf{X}))$. *Specifically,*

$$P(\mathbf{Z}|\mathbf{C}=c,\mathbf{X}) = (2\pi)^{-D/2}\det(\Sigma^c)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{Z}-\mu^c)^{\mathsf{T}}(\Sigma^c)^{-1}(\mathbf{Z}-\mu^c)\right),$$

where $\mathbf{Z} \in \mathbb{R}^D$. *If* $\Sigma^c(\mathbf{X})$ *is diagonal for all c, we have*

$$l_c = \sum_{i=1}^{D}[\mathbb{E}(Z_i|do^c(Z_{-i}),\mathbf{X}) - \mathbb{E}(Z_i|\mathbf{X})] = 0. \tag{7}$$

From Theorem 3.1, we see that for each $\mathbf{C} = c$, enforcing the latent variable $\mathbf{Z}$ to be statistically independent minimizes l_c. Taking the whole $\mathbf{C}$ set into consideration, $P(\mathbf{Z}|\mathbf{X})$ subjects to a mixture of Gaussian distribution where each centroid is inferred from observational data under a specific realization of the confounder $\mathbf{C}$:

$$P(\mathbf{Z}|\mathbf{X}) = \sum_{c\in C}\pi_c\mathcal{N}(\mu^c(\mathbf{X}),\Sigma^c(\mathbf{X})). \tag{8}$$

The mixing coefficient $\pi_c = P(\mathbf{C} = c|\mathbf{X})$ reads the probability of occurrence of $\mathbf{C} = c$. Nevertheless, such a hard assignment of coefficient varies with observable dataset and cannot accommodate scenarios in which the label set $\mathbf{C}$ does not exist. In this paper, we parameterize π_c as a Gaussian distribution for a soft assignment of samples: $\pi_c \sim \mathcal{N}(\mu^{\pi_c}(\mathbf{X}), \Sigma^{\pi_c}(\mathbf{X}))$. Parameters $\mu^{\pi_c}(\mathbf{X})$ and $\Sigma^{\pi_c}(\mathbf{X})$ are learned to minimize the discrepancy with $P(\mathbf{C} = c|\mathbf{X})$.

In VAEs, the prior of the latent space $P(\mathbf{Z})$ is assumed to follow a Gaussian distribution with mean zero and identity variance: $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$. To avoid enforcing statistical independence of the overall latent spaces we learn, we only assume that the prior of $\mathbf{Z}$ to be a normal distribution with variance one for each subset of $\mathbf{C}$. Specifically, suppose that the latent variable $\mathbf{Z}$ for $\mathbf{X}$ under $\mathbf{C} = c$ follows a Gaussian distribution $\mathcal{N}(\mu^c(\mathbf{X}), \Sigma^c(\mathbf{X}))$, then the KL divergence in (4) regulates the distribution to $\mathcal{N}(\mu^c(\mathbf{X}), I)$.

By the Lagrangian multiplier method, the new loss function is

$$\mathcal{L} = \underbrace{-\mathbb{E}[\log P(\mathbf{X}|\mathbf{Z})]}_{\mathcal{L}_{rec}} + \underbrace{\mathbb{E}[\log P(\mathbf{C}|\pi_c(\mathbf{X}))]}_{\mathcal{L}_{cls}} + D_{KL}[P(\mathbf{Z}|\mathbf{X},\mathbf{C})||P(\mathbf{Z}|\mathbf{C})].$$

4. Experiments

In this section, we experimentally compare cdVAE with various baselines on synthetic and real-world datasets, and study properties of cdVAE through the ablation studies. Concretely, we aim to answer the following questions regarding the proposed model:

- ▷ **Q1**: How does it perform compared to the existing methods in the latent space?
- ▷ **Q2**: How does it perform in downstream tasks such as classification under distribution shifts?

4.1. Basic setup

Datasets. we evaluate cdVAE on three datasets: synthetic datasets 3dshape (Burgess & Kim, 2018) and Candle (Reddy et al., 2022), and real-world dataset CelebA (Liu et al., 2015).

Baselines and tasks. We compare the performance of cdVAE in the task of image generation and classification under distribution shift with the following architectures: (1) VAE-based methods (vanilla VAE (Kingma & Welling, 2013), β-VAE (Higgins et al., 2016), FactorVAE (Kim & Mnih, 2018)) that equate disentanglement as statistical independence, (2) Existing causal regulation methods, CAUSAL-REP (Wang & Jordan, 2021), (3) cVAE (Kingma & Welling, 2013) as it also applies the label information (4) GM-

Table 1: **Compare with baselines in image generation task on celebA and candle.** The reconstruction error indicates the end-to-end performance of the image generation task. D-score measures from a non-causal perspective and requires that the generation process is unconfounded. We expect that a good method that recovers causally disentangled factors should obtain poor (i.e., low) D-scores. IOSS, UC and CG are causal metrics that measure the level of disentanglement of a representation.

	CelebA			Candle				
Methods	**Recon** ↓	**D** ↑ (non-causal)	**IOSS** ↓ (causal)	**Recon** ↓	**D** ↑ (non-causal)	**UC** ↑ (causal)	**CG** ↑ (causal)	**IOSS** ↓ (causal)
VAE	0.33	0.11	0.78	.024	0.14	0.10	0.18	0.69
β-VAE	0.27	0.15	0.74	.017	**0.18**	0.11	0.24	0.54
FactorVAE	0.25	**0.17**	0.68	.014	0.15	0.13	0.26	0.51
CAUSAL-REP	0.29	0.16	0.34	.012	**0.18**	0.20	0.32	0.31
cVAE	0.32	0.14	0.64	.020	0.14	0.12	0.21	0.62
GMVAE	0.30	0.13	0.71	.018	0.12	0.09	0.16	0.71
cdVAE	**0.18**	0.12	**0.21**	**.008**	0.11	**0.35**	**0.54**	**0.16**

Methods	**MIC** ↑	**TIC** ↑	**IRS** ↑
VAE	21.9	12.1	0.82
β-VAE	22.1	12.4	0.85
FactorVAE	24.3	15.6	0.89
CAUSAL-REP	26.8	16.1	0.88
cVAE	22.4	12.4	0.84
GMVAE	23.2	12.8	0.81
cdVAE	**31.9**	**20.2**	**0.89**

Table 2: Compare how ground truth generative factors are recovered (MIC/TIC) and how disentangled they are in the latent space in classification on 3dshape dataset with shift severity = 0.5.

VAE (Dilokthanakul et al., 2016) as it also adopts a mixture of Gaussian model in a variational autoencoder framework. A detailed introduction of these metrics can be found in Appendix E.

4.2. Experimental results

(Q1) cdVAE significantly outperforms various baselines in end-to-end measurement.
We compare our cdVAE with baseline models in the image generation task on the CelebA, and Candle datasets. More details can be found in Appendix E.

As shown in Table 1, cdVAE are evaluated by several groups of metrics. Our method outperforms all baselines on these metrics except for the D score. Note the D score is only an effective measurement when there are no confounders in the observational dataset, resulting in a low D score with our model as expected.

(Q2.1) cdVAE are more robust under distribution shifts.
We conduct the task of shape classification on the 3dshape dataset with distribution shift and use the classification accuracy as a metric for out-of-distribution generalization. The level of distribution shift is measured by the shift severity.

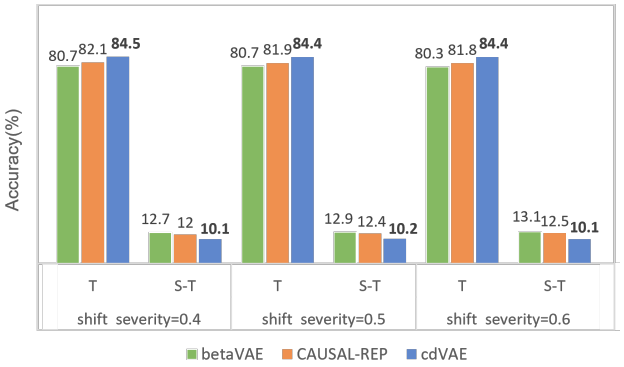

Figure 2: Compare cdVAE with β-Vae, CAUSAL-REP on classification under distribution shift. T represents accuracy on the target data, S-T represents the performance drop when the classifier trained on the source data is directly tested on the target data.

We train cdVAE, β-VAE, and CAUSAL-REP using images from the source set, with decoders being replaced by classifiers. The trained classifier is then tested on the targets set. We report the classification accuracy on the target set and the performance drop in Figure 2. We could observe that cdVAE is more robust than other baselines under distribution shift as it has the lowest performance drop and the highest target set accuracy.

(Q2.2) cdVAE better discovers the ground truth generative factors.
In the task of shape classification, we further examine how well the learned representations approximate the ground truth generative labels and to what extent they are causally disentangled. Table 2 shows that cdVAE outperforms all baselines in approximating the generative factors and the disentanglement.

References

Ahuja, K., Hartford, J. S., and Bengio, Y. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35: 15516–15528, 2022.

Brehmer, J., De Haan, P., Lippe, P., and Cohen, T. Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*, 2022.

Burgess, C. and Kim, H. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Chen, J. and Batmanghelich, K. Weakly supervised disentanglement by pairwise similarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3495–3502, 2020.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Horan, D., Richardson, E., and Weiss, Y. When is unsupervised disentanglement possible? *Advances in Neural Information Processing Systems*, 34:5150–5161, 2021.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Liu, X., Lu, H., Yuan, J., and Li, X. Cat: Causal audio transformer for audio classification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pp. 4402–4412. PMLR, 2019.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Pearl, J. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

Reddy, A. G., Balasubramanian, V. N., et al. On causally disentangled representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8089–8097, 2022.

Soch, J., of Statistical Proofs, T. B., Faulkenberry, T. J., Petrykowski, K., and Allefeld, C. StatProofBook/StatProofBook.github.io: StatProofBook 2020, December 2020. URL `https://doi.org/10.5281/zenodo.4305950`.

Suter, R., Miladinovic, D., Schölkopf, B., and Bauer, S. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pp. 6056–6065. PMLR, 2019.

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

Wang, Y., Liang, D., Charlin, L., and Blei, D. M. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 426–431, 2020.

Xu, C., Liu, C., Sun, X., Yang, S., Wang, Y., Wang, C., and Fu, Y. Patchmix augmentation to identify causal features in few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

A. Related Work

Disentangled Representations The pursuit for disentangled representation can be dated to the surge of representation learning and is always closely associated with the generative process in modern machine learning, following the intuition that each dimension should encode different features. (Chen et al., 2016) attempts to control the underlying factors by maximizing the mutual information between the images and the latent representations. (Eastwood & Williams, 2018) propose a quantitative metric with the information theory. They evaluate the disentanglement, completeness, and informativeness by fitting linear models and measuring the deviation from the ideal mapping. (Higgins et al., 2016; Kim & Mnih, 2018; Chen et al., 2018; Mathieu et al., 2019) encourage statistical independence by penalizing the Kullback-Leibler divergence (KL) term in the VAE objective. However, the non-causal definitions of disentanglement fail to consider the cases where correlated features in the observational dataset can be disentangled in the generative process. Such a challenge is well-approached through a line of research from the causal perspective.

Causal Generative Process. Causal methods are widely used for eliminating spurious features in various domains and improving understandable modelling behaviours(Wang et al., 2020; Xu et al., 2022; Liu et al., 2023). It is not until (Suter et al., 2019) that it was introduced for a strict characterization of the generative process. (Suter et al., 2019) first provided a rigorous definition of a causal generative process and the definition of disentangled causal representation as the non-existence of causal relationships between two variables, i.e., the intervention on one variable does not alter the distribution of the others. The authors further introduce *interventional robustness* as an evaluation metric and show its advantage on multiple benchmarks. (Reddy et al., 2022) follow the path of (Suter et al., 2019) and further propose two evaluation metrics and the Candle dataset. The confounded assumption allows for correlation in the latent space without tempering with the disentanglement in the data generative. Despite effective evaluation tools, there is still a missing piece on how to infer a set of causally disentangled features. Using the proposed evaluation metric as regulation, the model implicitly assumes unconfoundedness and it falls back to finding statistical independence in the latent space. The problem of unrealistic unconfoundedness assumption is identified by (Wang & Jordan, 2021). They assume that confounders exist but they are unobservable. They further propose an evaluation metric considering the existence of confounders, that causally disentangled latent variables have independent support measured by the IOSS score. Similar to the evaluation metrics introduced in (Suter et al., 2019; Reddy et al., 2022), IOSS is also a necessary condition of the causal disentanglement. More importantly, as in previous work focusing on obtaining statistical independence, such a regulation suffers from the identifiability issue.

Weak Supervision for Inductive Bias. The identifiability issue in unsupervised disentangled representation learning is first identified in (Locatello et al., 2019). Specifically, they show from the theory that such a learning task is impossible without inductive biases on both the models and the data. Naturally, a series of weak-supervised or semi-supervised methods (Chen & Batmanghelich, 2020; Ahuja et al., 2022; Brehmer et al., 2022) are proposed with a learning objective of statistical independence or alignment. In this paper, we take a step further for the confounding assumption, assuming that the confounders are observable with proper inductive bias so that the latent representation can be better identified. We, similarly, adopt partial labels of the dataset as the supervision signal. We treat the labels as a source of possible confounders and allow the learning of correlated but causally disentangled latent generative factors to be learned.

B. Preliminaries

In this section, we introduce basic concepts in causal inference and then show how to evaluate the causal relationship among variables in latent space.

Causal graph through DAG. The causal relationship among variables can be reflected by a Directed Acyclic Graph (DAG). Each (potentially high-dimensional) variable is denoted as a node. The directed edges indicate the causal relationships and always point from parents to children.

Intervention and do-operator. *Intervention* is one of the fundamental concepts in causal inference. When we intervene on a variable, we set the value of a variable and cut all incoming causal arrows since its value is thereby determined only by the intervention (Pearl, 2012). The intervention is mathematically represented by the *do-operator* $do(\cdot)$. Let Z_1 and Z_2 be two variables, $P(Z_2|do(Z_1 = z_1))$ characterizes an interventional distribution and reflects how a change of Z_1 affects the distribution of Z_2. The do-operation and the interventional distribution should be estimated on the interventional dataset. However, in practice, the true distribution of the data is unavailable but an observational subset. As a result, we estimate the interventional distribution from the observational set following *do-calculus* introduced by (Pearl, 2009).

Do-calculus consists of three rules that help with identifying causal effects.

Rule B.1 (Insertion/deletion of observations)**.**

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{9}$$

Rule B.2 (Action/observation exchange)**.**

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{10}$$

Rule B.3 (Insertion/deletion of actions)**.**

$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{XZ(W)}}} \tag{11}$$

where $G_{\overline{X}}$ is the graph with all incoming edges to X being removed, $G_{\underline{W}}$ is the graph with all outcoming edges to W being removed, and $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node.

Intuitively, Rule B.1 states when an observant can be omitted in estimating the interventional distribution, Rule B.2 illustrates under what condition, the interventional distribution can be estimated using the observational dataset, and Rule B.3 decides when we can ignore an intervention.

Confounders bring in spurious correlation. Although the detailed definitions vary from literature, a confounder usually refers to a common cause (i.e., a common parent in the causal graph) of multiple variables and it brings in a certain level of correlation among these variables. Consequently, to estimate the causal effect of one variable on the other from the observational dataset, we have to eliminate the correlation introduced by the confounders. For example, when analyzing the causal relationship between the number of heat strokes and the rate of ice cream consumption, we may find that there is a correlation. However, the temperature is a confounder in the situation. If we eliminate this spurious correlation by conditioning on the temperature, heat stroke is not correlated to the rate of ice cream consumption.

Evaluation of interventional distribution. We specifically consider the case where there are variables Z_1 and Z_2, and we analyze how the existence of parental nodes affects the estimation of $P(Z_2|do(Z_1 = z_1))$ on the observational distribution.

Proposition B.4. *Let Z_1 and Z_2 be two random variables, $\mathbf{C}^*$ be the ground truth confounder set. If $\mathbf{C}$ is a superset of or is equivalent to $\mathbf{C}^*$, i.e., $\mathbf{C}^* \subseteq \mathbf{C}$, with c being a realization of $\mathbf{C}$, we have*

$$P(Z_2|do(Z_1)) = \sum_{c \in \mathbf{C}} P(Z_2|Z_1, \mathbf{C} = c)P(\mathbf{C} = c) \tag{12}$$

if no $C \in \mathbf{C}$ is a descendent of $\mathbf{Z}$.

The proof can be found in Appendix D.1.

Intuitively speaking, Proposition B.4 states that to accurately estimate the causal relationship between variables, we have to eliminate the spurious correlation by conditioning on confounders. In addition, conditioning on additional variables will not affect the estimation if they are not decedents of $\mathbf{Z}$.

C. Relationship between $\mathbf{C}$ and $\mathbf{C}^*$

The generative factors $\mathbf{G}$ are not identifiable without proper inductive bias of $\mathbf{C}^*$ (Locatello et al., 2019). Previous works on discovering the generative factors either obtain disentanglement by enforcing statistical independence on latent variables (Higgins et al., 2016; Liu et al., 2015; Chen et al., 2016), or require that latent variables do not capture information of each other (Eastwood & Williams, 2018). Such a setting is equivalent to assuming unconfoundedness of the generative factors. It ignores the possibility that correlated latent variables can also be causally disentangled in the observational distribution, and hence is an assumption too restrictive. Fortunately, even though the ground truth generative factors are unobserved, domain expertise may inform a "reasonable" or "likely" inductive bias of the confounder from an accessible label set, denoted as $\mathbf{C}$. This $\mathbf{C}$ is used to account for all correlation among $\mathbf{Z}$.

Note here that we do not assume the accessible label set $\mathbf{C}$ equals to $\mathbf{C}^*$ (but hope that it is close to $\mathbf{C}^*$). The relationship between $\mathbf{C}^*$ and label set $\mathbf{C}$ must fall into one of the following scenarios, despite the immeasurability of their exact relationships.

Case 1 The label set contains no information about the confounders, i.e., $\mathbf{C} = \emptyset$

Case 2 The label set contains partial information about the confounders, i.e.,$\mathbf{C} \subset \mathbf{C}^*$,

Case 3 The label set contains all information about the confounders, i.e.,$\mathbf{C} = \mathbf{C}^*$

One may argue that $\mathbf{C}$ may contain information irrelevant to the ground truth confounders. We show in Appendix D that in the confounded generative process described in this paper, irrelevant information in $\mathbf{C}$ does not affect the evaluation of the interventional distribution, and therefore can be ignored. We only take into consideration how much information in $\mathbf{C}^*$ is captured here without loss of generality.

According to Proposition B.4, in case 3, we can estimate $P(Z_i|do(Z_{-i} = z_{-i}, \mathbf{X})$ on the observational set with inductive bias from $\mathbf{C}$. However, in the rest of the cases, the equation does not hold. Therefore, we introduce an operator $do^c(\cdot)$ to estimate the interventional distribution on the observational set under inductive $\mathbf{C}$. The framework that applies $do^c(\cdot)$ is named C-Disentanglement, shorten for *Confounded Disentanglement*.

D. Proofs

D.1. Proof of Proposition B.4

Proposition D.1. *Let Z_1 and Z_2 be two random variables, $\mathbf{C}^*$ be the ground truth confounder set. If $\mathbf{C}$ is a superset of or is equivalent to $\mathbf{C}^*$, i.e., $\mathbf{C}^* \subseteq \mathbf{C}$, with c being a realization of $\mathbf{C}$, we have*

$$P(Z_2|do(Z_1)) = \sum_{c \in \mathbf{C}} P(Z_2|Z_1, \mathbf{C} = c)P(\mathbf{C} = c) \tag{13}$$

if no $C \in \mathbf{C}$ is a descendent of $\mathbf{Z}$.

Proof.

$$\begin{aligned} P(Z_2|do(Z_1)) &= P(Z_2|do(Z_1), \mathbf{C})P(\mathbf{C}|do(Z_1)) \\ P(Z_2|do(Z_1), \mathbf{C}) &\overset{\text{Rule B.2}}{=\!=\!=} P(Z_2|Z_1, \mathbf{C}) \\ P(\mathbf{C}|do(Z_1)) &\overset{\text{Rule B.3}}{=\!=\!=} P(\mathbf{C}) \\ P(Z_2|do(Z_1)) &= \sum_{c \in \mathbf{C}} P(Z_2|Z_1, \mathbf{C} = c)P(\mathbf{C} = c) \end{aligned}$$

□

D.2. Proof of Theorem 3.1

Theorem D.2. *Suppose that the latent variable $\mathbf{Z}$ on dataset $\mathbf{X}$ given $\mathbf{C} = c$ is Gaussian $\mathcal{N}(\mu^c(\mathbf{X}), \Sigma^c(\mathbf{X}))$. Specifically,*

$$P(\mathbf{Z}|\mathbf{C} = c, \mathbf{X}) = (2\pi)^{-D/2} \det(\Sigma^c)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{Z} - \mu^c)^\intercal (\Sigma^c)^{-1}(\mathbf{Z} - \mu^c)\right),$$

where $\mathbf{Z} \in \mathbb{R}^D$. If $\Sigma^c(\mathbf{X})$ is diagonal for all c, we have

$$l_c = \sum_{i=1}^{D} [\mathbb{E}(Z_i|do^c(Z_{-i}), \mathbf{X}) - \mathbb{E}(Z_i|\mathbf{X})] = 0. \tag{14}$$

Proof. We suppose that

$$P(\mathbf{Z}|\mathbf{C} = c, \mathbf{X}) = (2\pi)^{-D/2} \det(\Sigma^c)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{Z} - \mu^c)^\intercal (\Sigma^c)^{-1}(\mathbf{Z} - \mu^c)\right) \tag{15}$$

where we omit $\mathbf{X}$ for simplicity and D is the dimension of $\mathbf{Z}$ for any given c. By definition of l_c (Equation (6)) and proposition B.4,

$$l_c = \sum_{i=1}^{D} d\left(\mathbb{E}[Z_i|do^c(Z_{-i}), \mathbf{X}] - \mathbb{E}[Z_i|\mathbf{X}]\right) \tag{16}$$

$$= \sum_{i}^{D} d(E[Z_i|Z_{-i}, \mathbf{X}, C=c], E[Z_i|\mathbf{X}, C=c]) \tag{17}$$

$$= \sum_{i}^{D} d(E[Z_i^c|Z_{-i}^c], E[Z_i^c]) \tag{18}$$

where we denote $\mathbf{Z}^c = [\mathbf{Z}|\mathbf{X}, C=c]$ for simplicity. Notice that $\mathbf{Z}^c \sim \mathcal{N}(\mu^c, \Sigma^c) \in \mathbb{R}^D$, we therefore know that the conditional distribution of any subset vector Z_k^c, given the complement vector Z_j^c, is also a multivariate Gaussian distribution (Soch et al., 2020)

$$Z_k^c|Z_j^c \sim \mathcal{N}(\mu_{k|j}^c, \Sigma_{k|j}^c) \tag{19}$$

where

$$\mu_{k|j}^c = \mu_k^c + \Sigma_{k,j}^c(\Sigma_{j,j}^c)^{-1}(Z_j^c - \mu_j^c), \quad \Sigma_{k|j}^c = \Sigma_{k,k}^c - \Sigma_{k,j}^c(\Sigma_{j,j}^c)^{-1}\Sigma_{j,k}^c, \tag{20}$$

given that $\Sigma_{j,j}^c$ is nonsingular.

Hence we know that the first expectation in Equation (18) becomes

$$E[Z_i^c|Z_{-i}^c] = \mu_i^c + \Sigma_{i,-i}^c(\Sigma_{-i,-i}^c)^{-1}(Z_{-i}^c - \mu_{-i}^c) \tag{21}$$

assuming that $\Sigma_{-i,-i}^c$ is nonsingular. Since $\mathbb{E}[Z_i^c] = \mu_i^c$, the loss l_c can be written as

$$l_c = \sum_{i}^{D} d(\mu_i^c + \Sigma_{i,-i}^c(\Sigma_{-i,-i}^c)^{-1}(Z_{-i}^c - \mu_{-i}^c), \mu_i^c). \tag{22}$$

We assume further that Σ^c is a diagonal matrix. Therefore $\Sigma_{-i,-i}^c = \mathbf{0}$ is a zero row vector. Then

$$l_c = \sum_{i}^{D} d(\mu_i^c, \mu_i^c) = 0 \tag{23}$$

□

E. Experimental Details

E.1. Basic setup

Datasets. we evaluate cdVAE on three datasets: synthetic datasets 3dshape (Burgess & Kim, 2018) and Candle (Reddy et al., 2022), and real-world dataset CelebA (Liu et al., 2015). 3dshape is a dataset of 3D shapes generated from 6 ground-truth independent latent factors. These factors are floor color, wall color, object color, scale, shape, and orientation. Candle is a dataset generated using Blender, a free and open-source 3D CG suite that allows for manipulating the background and adding foreground elements that inherit the natural light of the background. It has floor hue, wall hue, object hue, scale, shape, and orientation as latent factors. We use 3dshape and Candle as synthetic datasets.

Baselines and tasks. We compare the performance of cdVAE in the task of image generation and classification under distribution shift with the following architectures: (1) VAE-based methods (vanilla VAE (Kingma & Welling, 2013), β-VAE (Higgins et al., 2016), FactorVAE (Kim & Mnih, 2018)) that equate disentanglement as statistical independence, (2) Existing causal regulation methods, CAUSAL-REP (Wang & Jordan, 2021), (3) cVAE (Kingma & Welling, 2013) as it also applies the label information (4) GMVAE (Dilokthanakul et al., 2016) as it also adopts a mixture of Gaussian model in a variational autoencoder framework.

Note that in this paper, we do not compare cdVAE with CausalVAE (Yang et al., 2021), despite the latter also obtaining "disentanglement" in the latent space. The main reason is that CausaslVAE aims to disentangle known ground truth generative factors in the latent space, which is fundamentally different from the learning objective in this paper.

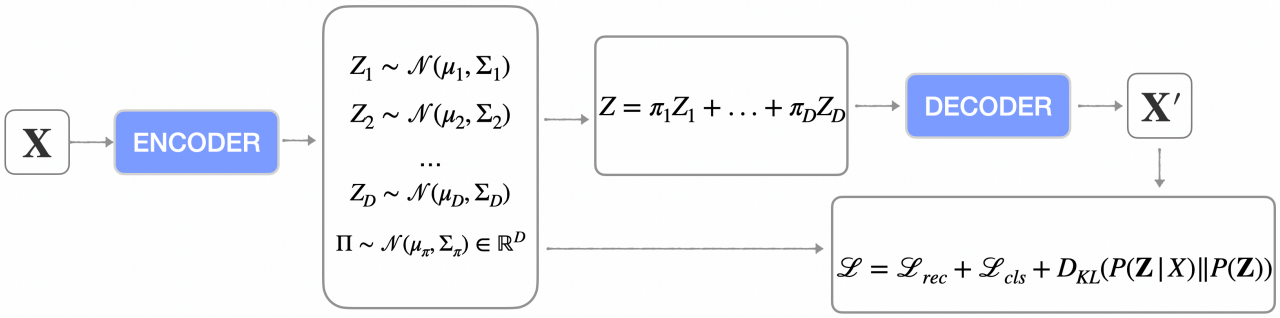

Figure 3: **Learning paradigm of cdVAE.** The input data $\mathbf{X}$ is partitioned by realizations of the confounder $\mathbf{C}$. We infer from each partition a Gaussian distribution and form a mixture of Gaussian model to characterize the distribution of $\mathbf{Z}$ on the observational distribution. We assume soft assignment of samples and further infer $\pi^c(\mathbf{X})$ that resembles $\mathbf{C}$.

Evaluation metrics. The experimental results are evaluated on (1) end-to-end evaluation metrics: the accuracy of classification under distribution shifts and reconstruction loss in image generation task. (2) how well the learned latent generative factors recover the ground truth one: Maximal Information Coefficient (MIC) and Total Information Coefficient (TIC) (Kinney & Atwal, 2014). (3) Existing disentanglement scores from a causal perspective: IRS (Suter et al., 2019), UC/CG (Reddy et al., 2022), IOSS (Wang & Jordan, 2021), and statistical perspective: D-Score (Eastwood & Williams, 2018). The higher these evaluation metrics, the better the model except for IOSS (the lower, the better).

The experiments are conducted on 4 NVIDIA GeForce RTX 2080Ti. In each experiment, we repeat 5 times with different seeds and report the averaged results. In all experiments, only partial information on the ground truth confounder is provided. Specifically, for example, the 3dshape dataset, we first make some predefined rules, such as " 70% cubes are red". Then we generate 700 red cubes and 300 cubes in other colors. The generation process naturally divides the dataset into different subgroups, and we can thus explicitly control how inductive bias is provided, i.e., the grouping. In the celebA dataset, since we do not have access to the ground truth generative factors, so we assume any label sets only contain partial information.

In the shape classification experiments on the 3dshape dataset with distribution shift, we use the classification accuracy as a metric for out-of-distribution generalization. Specifically, in the source distribution, we sample a certain percentage of images in which the object hue is correlated with the object shape (i.e., red objects are cubes). The rest of the images are evenly generated by disentangled factors, while in the target domain, all images are generated by disentangled factors. The proportion of highly-correlated data is denoted by *shift severity*. For example, shift severity = 0.4 means that 40% of training images are sampled under preset correlation between object hue and object shape.

E.2. Ablations studies

We further conduct ablation studies to show that providing inductive bias indeed improves the discovery of generative factors in the latent space. Concretely, we compare cdVAE with conditional VAE (cVAE) (Kingma & Welling, 2013) and GMVAE (Dilokthanakul et al., 2016). Compared with vanilla VAE, the method proposed uses label information to provide inductive bias to confounders for partitioning the observational dataset cVAE has the label information compared with vanilla VAE and GMVAE is a VAE modelled by a mixture of Gaussian. As shown in Table 1, cdVAE has universally better results, showing the necessity of introducing bias to factors discovered in the latent space.

F. Algorithms

We show the flow chart of cdVAE as in Figure 3.