CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models

Mengyue Yang[1] Furui Liu[1] Zhitang Chen[1] Xinwei Shen[2] Jianye Hao[1]
Jun Wang[3]

[1]Noah's Ark Lab, Huawei, Shenzhen, China
[2]The Hong Kong University of Science and Technology, Hong Kong, China
[3]University College London, London, United Kingdom

Introduction

Disentangled Representations

"A disentangled representation can be defined as one where single latent units are sensitive to changes in single (*independent*) generative (*ground-truth*) factors, while being relatively invariant to changes in other factors" [6].

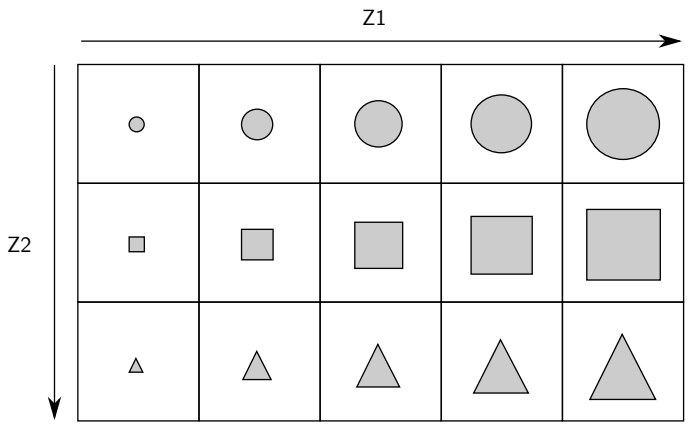


Figure 1: Toy example where each dimension of the latent space z controls a given generative factor: z_1 the size, z_2 the shape of the 2D objects.

VAE-based disentanglement learning in one slide

A significant amount of work in this area has been carried out for generative models like VAE [13, 21].

- Unsupervised disentanglement learning [8, 2, 11, 17, 5, 26]
- Impossibility result [18]
- Disentanglement through additionally observed variables [12, 7, 23, 20, 15, 1, 9, 19, 4, 3, 10]

What if ground-truth factors are not independent?

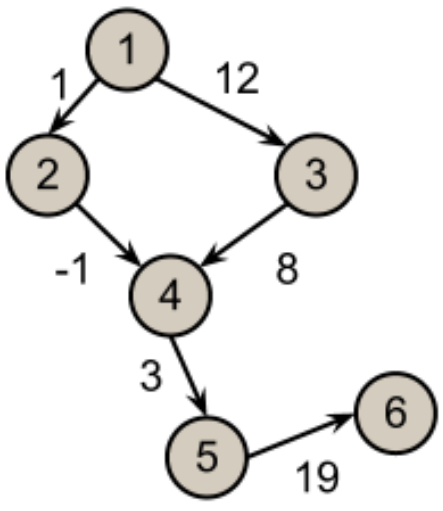
Disentangled representation learning builds on the assumption that the **ground-truth factors are independent**.

In many real world applications, they might be causally related.

Figure 2: A swinging pendulum: an illustrative example.

Causal Graphs in one slide

Causal graphs (also known as causal Bayesian networks or **DAGs**) are probabilistic graphical models used to encode assumptions about the data-generating process[1].



	1	2	3	4	5	6
1	0	1	12	0	0	0
2	0	0	0	-1	0	0
3	0	0	0	8	0	0
4	0	0	0	0	3	0
5	0	0	0	0	0	19
6	0	0	0	0	0	0

Figure 3: Example of causal graph (left) and its adjacency matrix (right).

[1] We focus on linear causal models.

CausalVAE

Model structure of CausalVAE

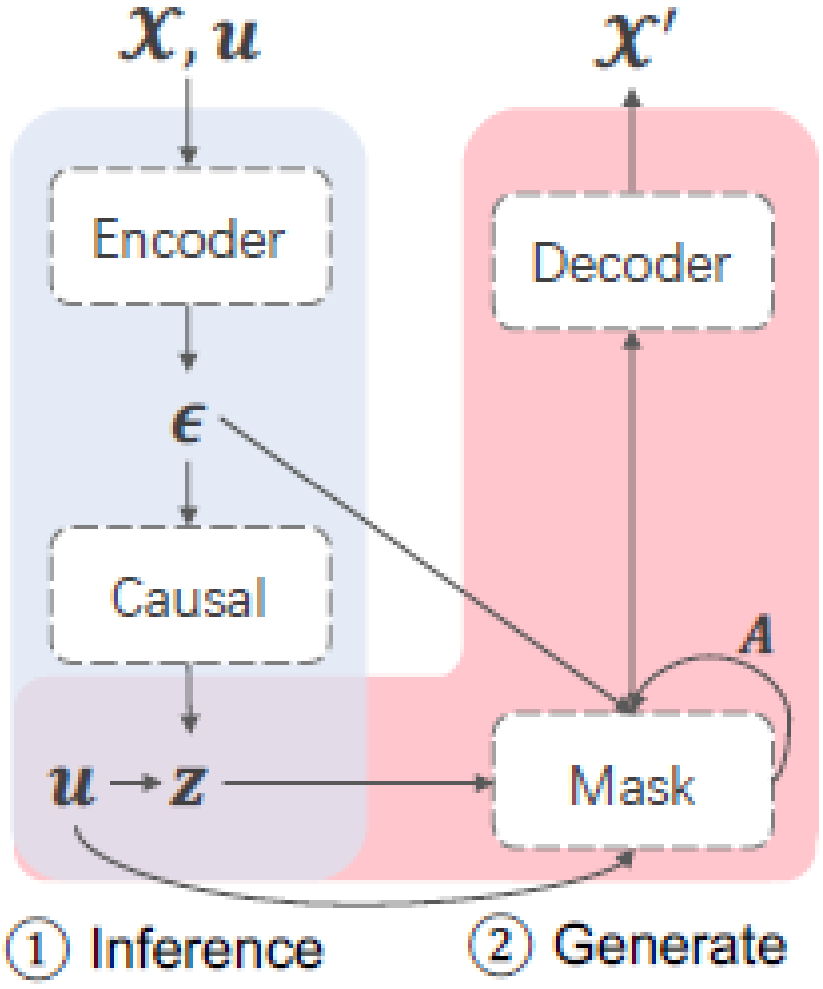


Figure 4: Model structure of CausalVAE

The Causal Layer

The key element of CausalVAE is the presence of a **Causal Layer**.

- The Causal Layer describes a Structural Causal Model (SCM) [22].
- The Causal Layer transforms **independent exogenous factors** ϵ into **causal endogenous factors** z corresponding to the ground-truth factors of interest.
  - The causal structure is defined by a DAG with adjacency matrix A.
- The Causal Layer is here restricted to implement a linear SCM:

$$z = A^T z + \epsilon = B^T \epsilon \tag{1}$$

A Probabilistic Generative Model for CausalVAE (1)

Let $\mathrm{x} \in \mathbb{R}^d$ and $\mathrm{u} \in \mathbb{R}^n$ be two observed variables[2]. Let $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be the latent exogenous independent variables and $\mathrm{z} \in \mathbb{R}^n$ be the latent endogenous variables. Equation (1) defines the relation between z and $\boldsymbol{\epsilon}$. Then, consider the following generative model:

$$p_{\boldsymbol{\theta}}(\mathrm{x}, \mathrm{z}, \boldsymbol{\epsilon}|\mathrm{u}) = p_{\boldsymbol{\theta}}(\mathrm{x}|\mathrm{z}, \boldsymbol{\epsilon}, \mathrm{u}) p_{\boldsymbol{\theta}}(\mathrm{z}, \boldsymbol{\epsilon}|\mathrm{u}), \tag{2}$$

where:

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

[2] In CausalVAE u is restricted to be the explicit label associated to each ground-truth factor.

A Probabilistic Generative Model for CausalVAE (2)

Let introduce the generative and inference models:

$$p_{\theta}(\mathrm{x}|\mathrm{z}, \epsilon, \mathrm{u}) = p_{\theta}(\mathrm{x}|\mathrm{z}), \tag{4}$$

$$q_{\phi}(\mathrm{z}, \epsilon|\mathrm{x}, \mathrm{u}) \tag{5}$$

corresponding to the following decoding and encoding processes:

$$\mathrm{x} = \mathrm{f}(\mathrm{z}) + \xi, \tag{6}$$

$$\mathrm{z} = \mathrm{h}(\mathrm{x}, \mathrm{u}) + \zeta, \tag{7}$$

where ξ and z are the vectors of independent noise with probability densities p_{ξ} and p_{ζ}[3].

[3]When ξ and z are infinitesimal, the encoder and decoder can be regarded as deterministic ones.

Learning Strategy

ELBO of CausalVAE

Given data set $\mathcal{D}$ with the empirical data distribution $q_{\mathcal{D}}(\mathrm{x}, \mathrm{u})$, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are learned by optimizing the following evidence lower bound (ELBO):

$$ELBO = \mathbb{E}_{q_{\mathcal{D}}}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathrm{z},\boldsymbol{\epsilon}|\mathrm{x},\mathrm{u})}[\log p_{\boldsymbol{\theta}}(\mathrm{x}|\mathrm{z}, \boldsymbol{\epsilon}, \mathrm{u})] - KL(q_{\boldsymbol{\phi}}(\mathrm{z}, \boldsymbol{\epsilon}|\mathrm{x}, \mathrm{u})||p_{\boldsymbol{\theta}}(\mathrm{z}, \boldsymbol{\epsilon}|\mathrm{u}))]. \tag{8}$$

Combining eqs. (1), (3) and (4) into eq. (8), by abuse of notation, we can write:

$$\begin{aligned} ELBO = \mathbb{E}_{q_{\mathcal{D}}}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathrm{z}|\mathrm{x},\mathrm{u})}[\log p_{\boldsymbol{\theta}}(\mathrm{x}|\mathrm{z})] \\ - KL(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathrm{x}, \mathrm{u})||p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon})) \\ - KL(q_{\boldsymbol{\phi}}(\mathrm{z}|\mathrm{x}, \mathrm{u})||p_{\boldsymbol{\theta}}(\mathrm{z}|\mathrm{u}))]. \end{aligned} \tag{9}$$

Learning Strategy

Learning the Causal Structure of Latent Codes

We need some additional constraints:

$$H(\mathrm{A}) = tr((\mathrm{I} + \mathrm{A} \circ \mathrm{A})^n) - n = 0, \tag{10}$$

$$l_u = \mathbb{E}_{q_\mathcal{D}} ||\mathrm{u} - \sigma \mathrm{A}^T \mathrm{u}||_2^2 \leq k_1. \tag{11}$$

Equation (10) is a DAG constraint [25].

Then:

$$\mathcal{L}_{\text{CausalVAE}} = ELBO - \alpha H(\mathrm{A}) - \beta l_u. \tag{12}$$

Experiments

Settings

Datasets Pendulum (pendulum angle, light angle, shadow length, shadow location), CelebA (gender, smile, eyes open, mouth open).

Metrics Maximal Information Coefficient (MIC) and Total Information Coefficient (TIC) [14].

Comparing Methods β-VAE [8], LadderVAE [24], DC-IGN [16].

MIC and TIC Results

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

Figure 5: The MIC and TIC between learned representation z and the label u.

The Learning Process of a Causal Matrix A

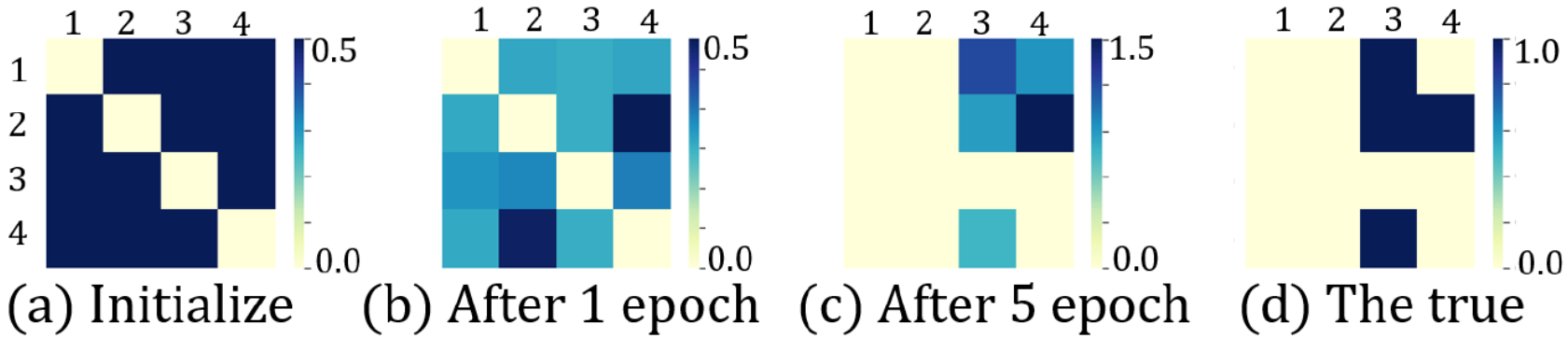


Figure 6: The learning process of causal matrix A for CelebA. (1=gender, 2=smile, 3=eyes open, 4=mouse open).

Intervention Experiments

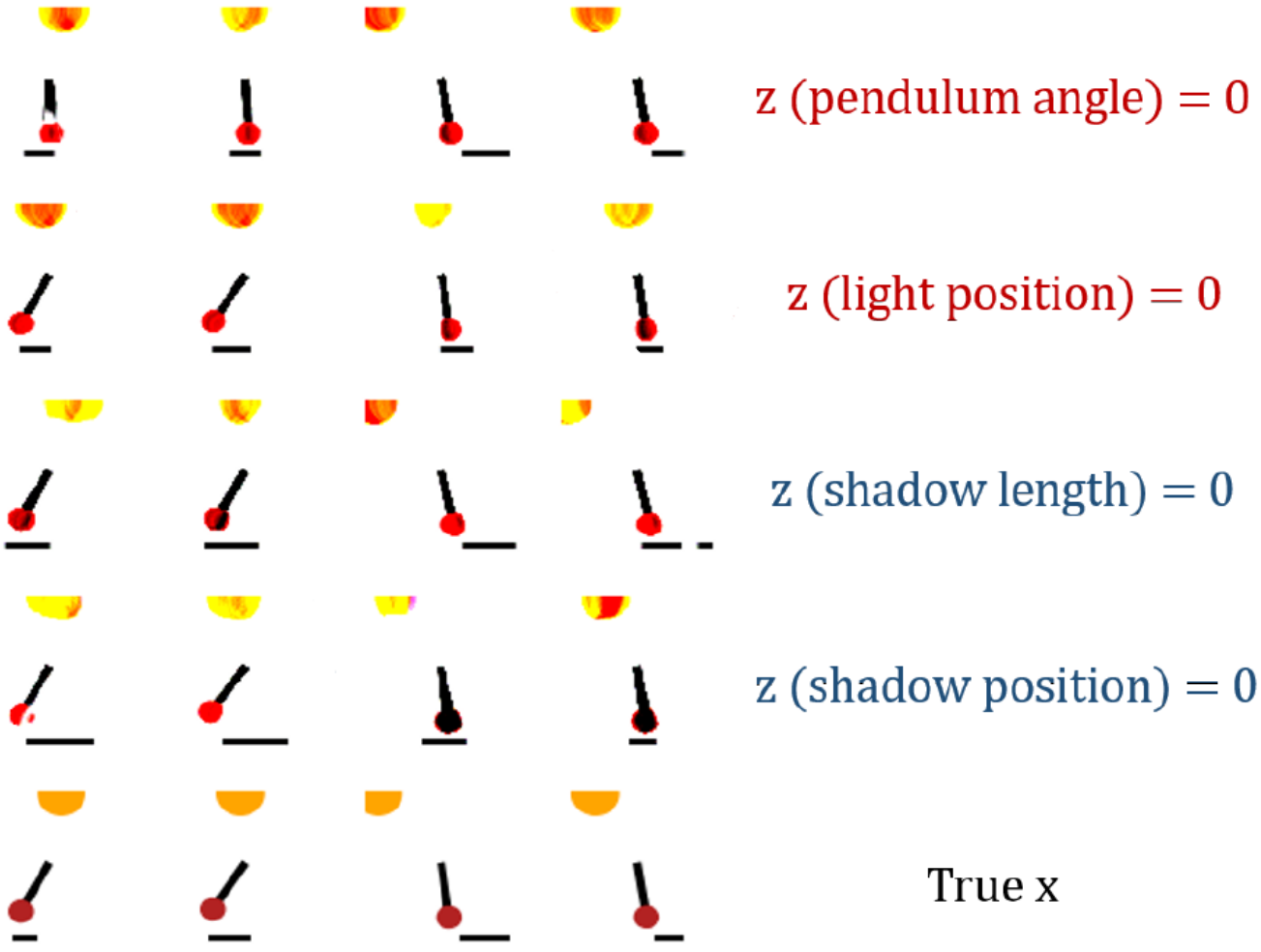


Figure 7: The results of Intervention experiments on the pendulum dataset.

Conclusions

Conclusions

According to the authors, CausalVAE is the **first work on causal disentanglement**.

- It allows to discover causal relationships among the ground-truth factors
  - Prior knowledge can be eventually incorporated into the Adjacency matrix
- It is **identifiable**
- It supports the so called **do-operation**
- It considers **linear** causal relationships
- It requires **full knowledge of the ground-truth factors**

References i

[1] D. Bouchacourt, R. Tomioka, and S. Nowozin.
Multi-level variational autoencoder: Learning disentangled representations from grouped observations.
In *Proc. of the 32nd AAAI Conf. on Artif. Intel.*, AAAI, 2018.

[2] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner.
Understanding disentangling in β-vae.
In *Proc. of the 30th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2017.

[3] J. Chen and K. Batmanghelich.
Robust ordinal vae: Employing noisy pairwise comparisons for disentanglement.
ArXiv, 2020.

References ii

[4] J. Chen and K. Batmanghelich.
Weakly supervised disentanglement by pairwise similarities.
In *Proc. of the 34th AAAI Conf. on Artif. Intel.*, AAAI, 2020.

[5] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud.
Isolating sources of disentanglement in variational autoencoders.
In *Proc. of the 31st Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2018.

[6] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel.
Infogan: Interpretable representation learning by information maximizing generative adversarial nets.
In *Proc. of the 29th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2016.

References iii

[7] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen.
Discovering hidden factors of variation in deep networks.
In *CoRR*, 2015.

[8] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner.
beta-vae: Learning basic visual concepts with a constrained variational framework.
In *Proc. of the 5th Int. Conf. on Learn. Repr.*, ICLR, 2017.

[9] H. Hosoya.
Group-based learning of disentangled representations with generalizability for novel contents.
In *Proc. of the 28th Int. Joint Conf. on Artif. Intel.*, IJCAI, 2019.

References iv

[10] I. Khemakhem, D. P. Kingma., R. P. Mont, and A. Hyvärinen.
Variational autoencoders and nonlinear ica: A unifying framework.
In *Proc. of the 23rd Int. Conf. on Artif. Intel. and Stat.*, AISTATS, 2020.

[11] H. Kim and A. Mnih.
Disentangling by factorising.
In *Proc. of the 35th Int. Conf. on Mach. Learn.*, ICML, 2018.

[12] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling.
Semi-supervised learning with deep generative models.
In *Proc. of the 27th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2014.

References v

[13] D. P. Kingma and M. Welling.
Auto-encoding variational bayes.
In *Proc. of the 2nd Int. Conf. on Learn. Repr.*, ICLR, 2014.

[14] J. B. Kinney and G. S. Atwal.
Equitability, mutual information, and the maximal information coefficient.
Proc. of the Nat. Acad. of Sc. of the Unit. Stat. of Americ., 111, 02 2014.

[15] J. Klys, J. Snell, and R. Zemel.
Learning latent subspaces in variational autoencoders.
In *Proc. of the 31st Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2018.

References vi

[16] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum.
Deep convolutional inverse graphics network.
In *Proc. of the 28th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2015.

[17] A. Kumar, P. Sattigeri, and A. Balakrishnan.
Variational inference of disentangled latent concepts from unlabeled observations.
In *Proc. of the 6th Int. Conf. on Learn. Repr.*, ICLR, 2018.

[18] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem.
Challenging common assumptions in the unsupervised learning of disentangled representations.
In *Proc. of the 36th Int. Conf. on Mach. Learn.*, ICML, 2019.

References vii

[19] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen.
Weakly-supervised disentanglement without compromises.
In *Proc. of the 37th Int. Conf. on Mach. Learn.*, ICML, 2020.

[20] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem.
Disentangling factors of variation using few labels.
In *Proc. of the 8th Int. Conf. on Learn. Repr.*, ICLR, 2020.

[21] D. J. Rezende, S. Mohamed, and D. Wiestra.
Stochastic backpropagation and approximate inference in deep generative models.
In *Proc. of the 31st Int. Conf. on Mach. Learn.*, ICML, 2014.

References viii

[22] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen.
A linear non-gaussian acyclic model for causal discovery.
JMLR, 7(72):2003–2030, 2006.

[23] N. Siddharth, B. Paige, J. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr.
Learning disentangled representations with semi-supervised deep generative models.
In *Proc. of the 30th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2017.

[24] C. K. Sø nderby, T. Raiko, L. Maalø e, S. r. K. Sø nderby, and O. Winther.
Ladder variational autoencoders.
In *Proc. of the 29th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2016.

[25] Y. Yu, J. Chen, T. Gao, and M. Yu.
Dag-gnn: Dag structure learning with graph neural networks.
In *Proc. of the 36th Int. Conf. on Mach. Learn.*, ICML, 2019.

[26] S. Zhao, J. Song, and S. Ermon.
Infovae: Balancing learning and inference in variational autoencoders.
In *Proc. of the 33rd AAAI Conf. on Artif. Intel.*, AAAI, 2019.