Do-Operation Guided Causal Representation Learning with Reduced Supervision Strength

Jiageng Zhu[1,2,3] **Hanchen Xie**[2,3] **Wael AbdAlmageed**[1,2,3]
[1] USC Ming Hsieh Department of Electrical and Computer Engineering
[2] USC Information Sciences Institute
[3] Visual Intelligence and Multimedia Analytics Laboratory
{jiagengz, hanchenx, wamageed}@isi.edu

Abstract

Causal representation learning has been proposed to encode relationships between factors presented in the high dimensional data. However, existing methods suffer from merely using a large amount of labeled data and ignore the fact that samples generated by the same causal mechanism follow the same causal relationships. In this paper, we seek to explore such information by leveraging *do-operation* for reducing supervision strength. We propose a framework which implements *do-operation* by swapping latent cause and effect factors encoded from a pair of inputs. Moreover, we also identify the inadequacy of existing causal representation metrics empirically and theoretically and introduce new metrics for better evaluation. Experiments conducted on both synthetic and real datasets demonstrate the superiorities of our method compared with state-of-the-art methods.

1 Introduction

Causal representation learning [14] has been proposed to extract causal relations from high dimension observations. To this end, CausalVAE [16] contains a causal layer and a mask layer as parts of deep neural network (DNN) architecture, and uses labels of generative factors to learn the causal relationship between different latent factors.

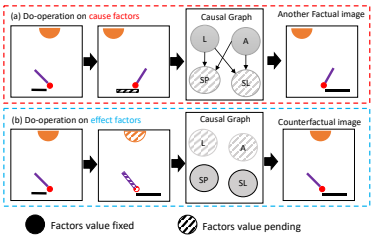


Figure 1: Do-operation *to cause and effect factors. Light position (L) and pendulum angle (A) are the cause of shadow position (SP) and shadow length (SL). Applying* do-operation *to cause factors will change the effect factors accordingly. Oppositely, applying* do-operation *to the effect factors will not affect the cause factors, and the original causal relationships from L and A to SP and SL are removed. Thus, a counterfactual sample will be created.*

However, training CausalVAE requires labels of all generative factors, which may still pose a strong assumption. For instance, all semantic causal factors need to be carefully annotated, which is either costly or hard to be identified in the first place. Further, since it relies on full supervision, CausalVAE limits the dimensionality of the latent representation to be the same as the number of generative factors and leaves no space for other *unknown* confounding factors which can be entangled with semantically meaningful latent factors and harm the performance. Moreover, CausalVAE incorporates ground-truth generative factors so that causal layer of CausalVAE can be trained separately from VAE. Thus, rather than extracting causal relations from high dimension observations, causal representation in CausalVAE is merely obtained through ground-truth generative factors, which is also used as a part of inputs during training.

To avoid the constraint of using fully supervised training, we utilize the *do-operation*, illustrated in Figure 1, to learn causal representation with reduced supervision. *Do-operation* [13] defines an

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

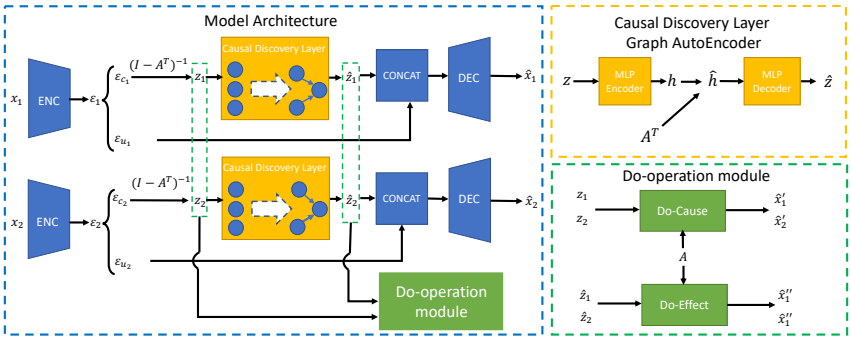


Figure 2: *Model structure. The input x is encoded to exogenous variable ε, which can be further splits into latent causal factors ε_c and unknown nuisance factors ε_u. ε_c is then mapped to endogenous variable z. The causal relationships are discovered and calculated through causal discovery layer. The unknown nuisance factors ε_u and causal representation $\hat{z}$ is then concatenated as the inputs of a decoder. A pair of inputs are used to introduce supervision signal. Two encoders and two decoders in model share same weights respectively.*

intervention that remove certain relationships in the causal graph and replace a factor with a constant. According to Pearl *et al.* [13], the causal effects can only propagate from cause factors to effect factors and not inversely. Thus, when *do-operation* is applied to cause factors, a new and factual sample will be generated. Conversely, when *do-operation* is applied to effect factors, the cause factors should be unaffected. Further, since *do-operation* changes the values of effect factors to constants, the newly generated sample can be counterfactual. When training the model, since the supervision strength is reduced to limited or even no labels, we use two latent representations encoded from a pair of inputs and apply *do-operation* via exchanging their latent factors with each other. By comparing the new reconstructions after *do-operation* with the original inputs, a supervision signal can be introduced.

CausalVAE [16] uses MIC and TIC [8] to evaluate the performance of causal representation learning. However, MIC and TIC only calculate mutual information between the latent representation and its corresponding ground truth generative factors. We argue, therefore, that MIC and TIC can only reflect the correctness of the marginal distribution of each factor itself, whereas no causal relationship between factors can be measured. Therefore, we propose new metrics for better evaluation.

Related work: Disentangled representation learning aims at attaining mutual independent latent factors [1] and Variational Autoencoder (VAE) [7] is the basic framework of most disentanglement methods, where the loss function is $L_{VAE}(x,z) = -\mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] + D_{KL}(q_\phi(z|x)||p(z))$. Other unsupervised VAEs including β-VAE [5] AnnealedVAE [2], LadderVAE [9] and β-TCVAE [3] are proposed by modifying L_{VAE}. Causal representation learning is the extension of disentangled representation learning. To achieve causal representation, CausalVAE [16], built upon iVAE [6], removes the requirement of prior knowledge of true causal graph by introducing the causal layer and mask layer into the model. However, all generative factor labels are required to train CausalVAE.

2 Method

Model architecture: We propose a new architecture, shown in Figure 2, as well as a training algorithm that greatly reduces the supervision strength via a *do-operation* module. We use x denotes an input image, $\varepsilon = [\varepsilon_c, \varepsilon_u]$ denotes exogenous latent factors which is further split into causal and unknown nuisance exogenous factors, z denotes latent causal factors, $\hat{z}$ denotes latent causal factors after causal discovery and $\hat{x}$ denotes reconstructed images.

In contrast with CausalVAE, our framework uses ε_u to encode unknown nuisance factors. Meanwhile, similar to CausalVAE, the exogenous factors ε_c are first transformed to endogenous latent factors z, and a causal discovery layer (CDL) propagates causal effects from parent factors to their child factors. We use a graph autoencoder (GAE) [12] as CDL, which learns nonlinear causal relationships and thus generalizes over NOTEARS [17] used in CausalVAE. The unknown nuisance latent factors ε_u are concatenated with the latent causal factors $\hat{z}$ as the input of a decoder. As discussed in [10], unsupervised learning can not identify expected latent representations so that supervision is necessary. To reduce supervision strength in CausalVAE and inspired by [11], we use a pair of inputs and

implement *do-operation* during training to utilize a weak supervision signal. The CDL, that applies causal effect from parent factors to child factors, is the key to implement *do-operation*, descried in Section 2, in order to decrease supervision strength. By using this new training strategy, we show in Section 4 that no label is needed during training on synthetic datasets and only a small amount of labels is needed on real datasets.

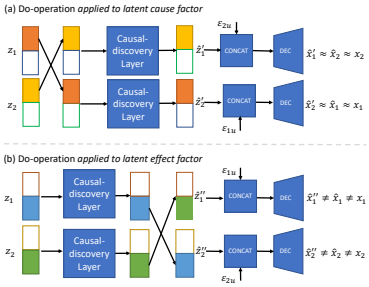


Figure 3: Do-operation *on cause factors encourage model to learn correct causal relationships, while* do-operation *on effect factors prevent model learning wrong causal relationship.*

Do-operation illustrates causal relationship: *Do-operation* [13] defines an intervention that deletes a specific relationship in the causal graph and replaces factors with constants. As shown in Figure 1, if *do-operation* is applied to cause factors, since the original causal graph stays unchanged, the effect factors will be affected according to their parent factors. Conversely, when *do-operation* is applied to effect factors, cause factors will not affect the value of effect factors. This process can be shown in Equations (1) and (2).

$$do(z_c^{(l)}) := z_c^{(m)}; \quad f([do(z_c^{(l)}), z_e^{(l)}]) = [z_c^{(m)}, z_e^{(m)}]; \quad (1)$$

$$do(z_e^{(l)}) := z_e^{(m)}; \quad f([z_c^{(l)}, do(z_e^{(l)})]) = [z_c^{(l)}, \ z_e^{(m)}]; \quad (2)$$

where f is the causal relationship function, and generative factors z are split into cause factors z_c and effect factors z_e. By assigning previous cause factors $z_c^{(l)}$ with new value $z_c^{(m)}$, effect factors z_e will change accordingly. Oppositely, if *do-operation* is applied to effect factors $z_e^{(l)}$ whose value is replaced by $z_e^{(m)}$, cause factors z_c should stay unchanged. Besides, The output of causal function f can be counterfactual since the original causal relationship has changed.

Do-operation on cause factors (Do-Cause): As shown in the Equation (1), if we apply *do-operation* to cause factors z_c, since z_c have no parent factors, the causal graph is unchanged and the value of effect factors z_e should change accordingly. To train our model, since no label or limited labels of generative factors are available, we use pairs of images as a weak supervision signal to encourage the model to learn causal representation. As illustrated in Figure 3a, except the regular propagation of inputs, after two endogenous latent factors z_1 and z_2 are encoded from a pair of inputs x_1 and x_2, we exchange the cause factors of two latent representations with each other to create two new latent representation $z_1' = [do(z_{1c}), z_{1e}]$ and $z_2' = [do(z_{2c}), z_{2e}]$, where a latent factor is cause or effect is determined by the learnable causal matrix A in CDL. As shown in Equation (3), two new representations z_1' and z_2' are fed into the CDL and then concatenated with their corresponding unknown nuisance factors ε_{u_1} and ε_{u_2} as inputs of the decoder.

$$\begin{aligned} z_1' &:= [do(z_{c_1}), z_{e_1}] = [z_{c_2}, z_{e_1}]; \quad \hat{z}_1' = f(z_1'); \quad \hat{x}_1' = Dec(\hat{z}_1', \varepsilon_{u_2}) \\ z_2' &:= [do(z_{c_2}), z_{e_2}] = [z_{c_1}, z_{e_2}]; \quad \hat{z}_2' = f(z_2'); \quad \hat{x}_2' = Dec(\hat{z}_2', \varepsilon_{u_1}) \end{aligned} \quad (3)$$

Recall that from Equation (1), the new outputs of CDL should be same with the original outputs of CDL, where $\hat{z}_1' = \hat{z}_2$ and $\hat{z}_2' = \hat{z}_1$, since *do-operation* on cause factors does not change causal graph, and the unchanged causal graph propagates causal relationships from cause factors to effect factors. Since the new latent causal representation $\hat{z}_2'$ and $\hat{z}_1'$ should be same with the original latent causal representation z_1 and z_2, their corresponding reconstructions $\hat{x}_1'$ and $\hat{x}_2'$ after the decoder should also be same with the the original inputs x_2, x_1. As shown in Equation (4), by comparing new reconstructions with the original inputs, the model is encouraged to learn the correct causal relationships, where d is distance function, such as binary cross entropy or mean square error.

$$L_{cause} = d(\hat{x}_1', x_2) + d(\hat{x}_2', x_1) \quad (4)$$

Do-operation on effect factors (Do-Effect): Compared with *do-operation* on the cause factors, since the causal graph will change when applying *do-operation* to the effect factors, the latent effect factors should be exchanged after the CDL in order to remove the effect of cause factors. The whole process of *do-operation* on the effect factors can be shown in Equation (5).

$$\begin{aligned} \hat{z}_1 &= f(z_1); \quad \hat{z}_1'' := [\hat{z}_{c_1}, do(\hat{z}_{e_1})] = [\hat{z}_{c_1}, \hat{z}_{e_2}]; \quad \hat{x}_1'' = Dec(\hat{z}_1'', \varepsilon_{u_1}) \\ \hat{z}_2 &= f(z_2); \quad \hat{z}_2'' := [\hat{z}_{c_2}, do(\hat{z}_{e_2})] = [\hat{z}_{c_2}, \hat{z}_{e_1}]; \quad \hat{x}_2'' = Dec(\hat{z}_2'', \varepsilon_{u_2}) \end{aligned} \quad (5)$$

Since *do-effect* changes the existing causal graph, the new latent representations $\hat{z}_1''$ and $\hat{z}_2''$ are not consistent with their corresponding latent representations z_1 and z_2. Thus, after decoder, the new reconstructions $\hat{x}_1''$ and $\hat{x}_2''$ will be different from their original inputs x_1 and x_2. Further, the new reconstructions are are actually counterfactual images as illustrated in Figure 1b. In practice, using MSE or BCE may lead to degenerated solution where $\hat{x}''$ are random noise. To solve this issue, we use a classifier C_w to distinguish factual images, including x_i, $\hat{x}_i$ and $\hat{x}_i'$, with counterfactual images $\hat{x}_i''$, where classifier and VAE are trained alternatively. The losses of training classifier and *do-operation* on the effect factors are shown in Equations (6) and (7) respectively.

$$L_{cla} = \mathbf{BCE}(C_w(x_i), \mathbf{1}) + \mathbf{BCE}(C_w(\hat{x}_i), \mathbf{1}) + \mathbf{BCE}(C_w(\hat{x}_i'), \mathbf{1}) + \mathbf{BCE}(C_w(\hat{x}_i''), \mathbf{0}) \quad (6)$$

$$L_{effect} = \mathbf{BCE}(C_w(\hat{x}_1''), \mathbf{0}) + \mathbf{BCE}(C_w(\hat{x}_2''), \mathbf{0}) \quad (7)$$

Training model with reduced supervision strength: As discussed in Section 1 and empirically proven in Section 4, our method only requires a small amount of supervision to train. For synthetic datasets, where actually no label are needed, the loss function is shown in Equation (8), where L_{VAE} is same with L_{VAE} shown in Section 1.

$$L_{no-label} = L_{VAE}(x, z) + \alpha L_{cause} + \beta L_{effect} + \gamma ||\hat{z} - z||_2^2 + h(A) \quad (8)$$

where α, β and γ are hyperparameters for regularizations. $||\hat{z}_i - z_i||_2^2$ is added to the loss since the outputs of CDL should align with their inputs. $h(A)$ is an acyclicity constraint for the causal graph A. In our implementation, we use $h(A) = tr(e^{A \odot A}) - d$ as proposed in [17].

If some labels of generative factors are available, similar to CausalVAE [16], we utilize them by adding label constrains to Equation (8) which leads to Equation (9), where f is CDL.

$$L_{semi} = L_{no-label} + ||u - f(u)||_2^2 + D_{KL}(q_\phi(z|x, u)||p(z|u)) \quad (9)$$

3 Evaluation Metrics For Causal Representation Learning

Maximum Information Coefficient (MIC) and Total Information Coefficient (TIC) [8] have originally been proposed as general purpose metrics to measure correlation between two random variables. Both metrics range from 0 to 1 and the higher value indicates better performance.

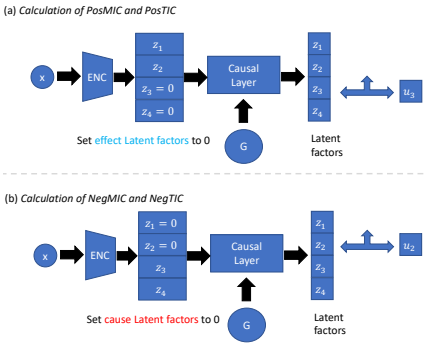


Figure 4: *One simple example of calculating new metrics.*

CausalVAE [16] suggested using MIC and TIC for evaluating causal representation learning, despite the following inadequacy. In CausalVAE, MIC and TIC first calculate the information relevance between every ground truth labels and their corresponding learned latent factors. Then, the means of MIC and TIC for every factors are used as the final metrics values. However, MIC and TIC only measure correlations between a latent factor and its corresponding generative factor, and can not evaluate the correctness of relationships between cause and effect factors. Therefore, we argue that MIC and TIC are not suitable for evaluating causal representation learning where the goal is to learn the correct causal relationships between cause and effect factors. An intuitive example for illustrating the deficiency of MIC and TIC can be found in Appendix.

To address this issue, we propose four new metrics: PosMIC, PosTIC, NegMIC and NegTIC. PosMIC and PosTIC are used to evaluate the causal relation correctness between latent factors, where higher value are expected. NegMIC and NegTIC are used to evaluate the falseness of causal relation discovery among latent factors, where lower value are expected. Additionally, to fully characterize the performance of causal representation learning using a single metric, we propose using the harmonic mean of the new metrics, i.e. F_1^{MIC} and F_1^{TIC}. We will first describe how the proposed new metrics are calculated and then discuss their adequacy over the metrics used in CausalVAE.

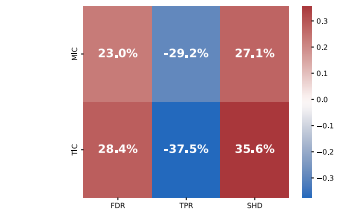
Calculating PosMIC, PosTIC, NegMIC and NegTIC: As illustrated in Figure 4, to calculate PosMIC and PosTIC, given ground truth causal graph G, we first set the latent effect factors (z_3 and z_4 in Figure 4) to 0. If the causal layer learns the correct relationship between the latent cause factors and the latent effect factors, z_3 and z_4 values are determined by the cause factors z_1 and z_2. Then, we separately calculate the MIC/TIC values of the latent effect factors and their corresponding

Table 1: *Causal representation metrics tested on Pendulum and Flow.*

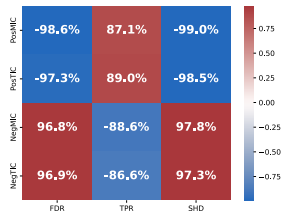
Models	Pendulum								Flow							
	MIC ↑	TIC ↑	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑	MIC ↑	TIC ↑	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑
Fully Supervised learning methods (all labels are used)																
CausalVAE	**95.1±2.1**	**81.6±1.9**	53.0±4.5	43.4±3.7	46.6±3.9	37.0±4.2	53.2±3.6	51.4±3.2	72.1 ±1.3	56.4 ±1.6	45.1 ±4.8	36.7 ±4.2	43.3 ±5.1	33.7 ±3.2	50.2 ±4.4	47.3 ±3.7
ConditionVAE	93.8±3.3	80.5±1.4	36.5±3.0	27.8±3.2	34.6±4.2	25.7 ±3.6	46.9 ±4.7	40.5 ±3.5	**75.5 ±2.3**	**56.5 ±1.8**	28.6 ±3.2	21.3 ±3.1	27.2 ±2.8	20.6 ±2.7	41.1 ±5.1	33.6 ±4.0
Unsupervised Learning methods (no label is used)																
CausalVAE(unsup)	21.2 ±1.4	12.0 ±1.0	20.5 ±2.6	11.8 ±2.7	23.3 ±3.2	14.7 ±1.9	32.4 ±3.4	20.7 ±3.1	20.5 ±4.7	11.8 ±2.6	22.8 ±2.7	12.5 ±1.4	21.5 ±2.4	12.0 ±1.9	35.3 ±5.6	21.9 ±4.7
BetaVAE	22.6 ±4.6	12.5 ±2.2	21.2 ±2.7	12.7 ±2.9	23.7 ±3.1	12.6 ±1.9	33.2 ±3.3	22.2 ±2.7	23.6 ±3.2	12.5 ±0.6	23.6 ±3.6	12.5 ±1.9	22.1 ±2.5	11.4 ±1.9	36.2 ±4.9	21.9 ±4.2
LadderVAE	22.4 ±3.1	12.8 ±1.2	15.2 ±1.9	8.6 ±1.0	**14.2 ±1.7**	**7.9 ±0.9**	25.8 ±3.0	15.7 ±2.8	34.3 ±4.3	24.4 ±1.5	16.2 ±1.8	10.5 ±1.0	**13.3 ±1.2**	**6.9 ±0.6**	27.3 ± 3.2	18.9 ±2.8
Reduced supervision method (no label is used; supervision source is image pairing)																
Our method	86.6 ±7.9	74.5 ±5.1	**54.1 ±4.5**	**44.0 ±4.2**	40.2 ±3.9	31.6 ±3.2	**56.8 ±5.2**	**53.6 ±4.3**	65.5 ±6.6	56.7 ±4.9	**50.7 ±4.7**	**41.3 ±4.2**	36.8 ±3.8	27.2 ±3.0	**56.3 ±5.9**	**52.7 ±4.9**

generative factors. Finally, the means of the MIC/TIC of all latent effect factors values are taken to be the PosMIC and PosTIC values. NegMIC and NegTIC are calculated in the opposite way, where the latent cause factors are set to 0, and the final MIC/TIC values are calculated between the latent cause factors after the causal layer and their corresponding generative factors. Ideally, the causal relationship should unidirectionally propagates from cause to effect, not in the opposite direction. Thus, the lower NegMIC and NegTIC indicate better performance of causal representation learning. To better compare different models and fully characterize the performance of causal representation learning, we consider Pos and Neg metrics together by calculating the harmonic mean: $F_1^{MIC} = 2 * \frac{PosMIC \cdot (1-NegMIC)}{PosMIC + (1-NegMIC)}$. F_1^{TIC} of PosTIC and NegTIC is calculated similarly.

Adequacy of proposed metrics: By conducting experiments on the Pendulum dataset, introduced in Section 4, we empirically show the advantage of the new metrics by proving that MIC and TIC fail to distinguish between models with correct and wrong causal graphs. We initialize causal graphs A of the CausalVAEs with different causal graphs and stop the gradient of elements if they are initialized with zero, such that CausalVAEs are created with various correctness levels of the causal graphs. If a causal graph A is initialized identical to the correct causal graph, the performance of that CausalVAE is expected to be optimal since the correct causal relationship is obtained by initialization. Conversely, the performance of wrong causal graph initialized CausalVAE is expected to be poor. After training, we calculate correlations among metrics used for causal representation learning and rubrics used in the causal discovery research area: True Positive Rate (TPR), False Discovery Rate (FDR), and Structural Hamming Distance (SHD). TPR and FDR calculate the rate of discovering correct and wrong causal relations, respectively. SHD is the minimum number of modifications to correct a causal graph. As shown in Figure 5, MIC and TIC have a low correlation with TPR, FDR, and SHD. In contrast, our proposed new metrics PosMIC, PosTIC, NegMIC, and NegTIC have significant higher correlation with three rubrics used in causal inference. PosMIC and PosTIC are more positively correlated with TPR, and NegMIC and NegTIC are positively correlated with FDR and SHD.



(a) *Correlation between MIC/TIC and FDR/TPR/SHD*

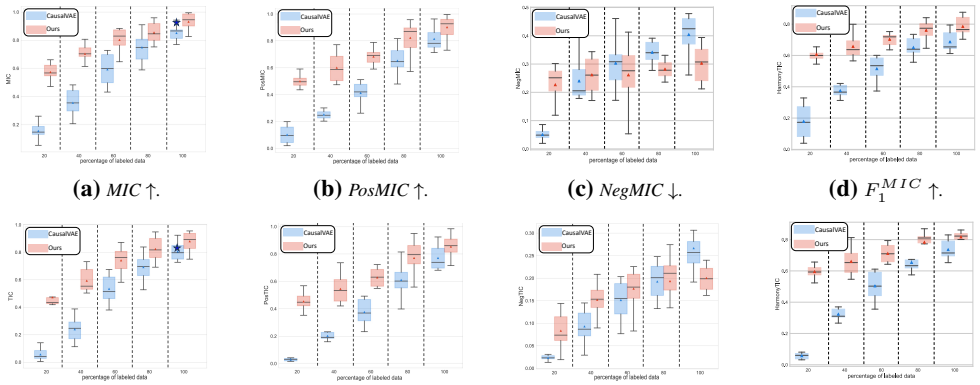


(b) *Correlation between new metrics and FDR/TPR/SHD*

Figure 5: *Correlation of different metrics on Pendulum dataset. MIC and TIC show low correlation with rubrics for causal discovery. Contrarily, our proposed metrics shows high and expected correlation with those rubrics.*

4 Experimental Evaluation

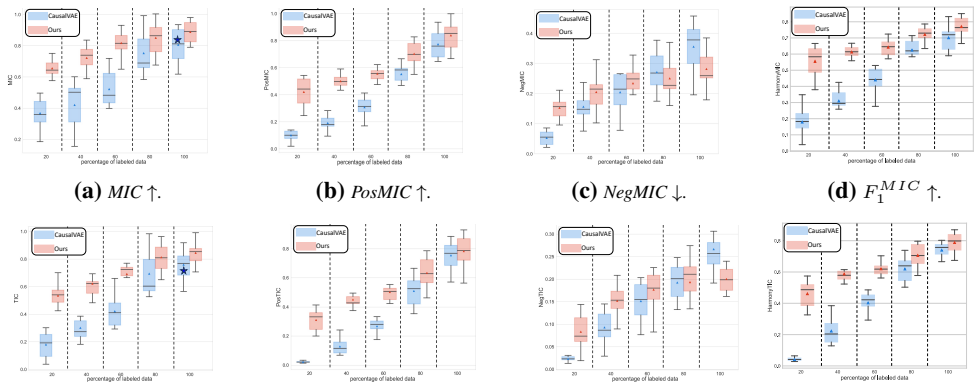
Datasets: Following [16], we use two synthetic datasets and two real world datasets. **Pendulum** focuses on pendulum angle, light angle, shadow location and shadow length, and **Flow** focuses on ball size, water height, hole and water flow. **CelebA(SMILE)** focuses on gender, smile, eyes open and mouth open, and **CelebA(BEARD)** focuses on age, gender and beardedness and baldness. We refer readers to [16] and Appendix for more details. Besides using MIC and TIC as evaluation metrics, we also use our new metrics for better evaluating causal representation learning.



(a) *MIC* ↑. **(b)** *PosMIC* ↑. **(c)** *NegMIC* ↓. **(d)** F_1^{MIC} ↑.

(e) *TIC* ↑. **(f)** *PosTIC* ↑. **(g)** *NegTIC* ↓. **(h)** F_1^{TIC} ↑.

Figure 6: *Box plots of metrics tested on CelebA(BEARD). Our method outperforms CausalVAE under various supervision strengths, where the advantage of our method is better revealed with weaker supervision strength. All experiments results are reproduced by us, except the blue star is the mean value reported in [16].*



(a) *MIC* ↑. **(b)** *PosMIC* ↑. **(c)** *NegMIC* ↓. **(d)** F_1^{MIC} ↑.

(e) *TIC* ↑. **(f)** *PosTIC* ↑. **(g)** *NegTIC* ↓. **(h)** F_1^{TIC} ↑.

Figure 7: *Box plots of metrics tested on CelebA(SMILE). Our method outperforms CausalVAE under various supervision strengths, where the advantage of our method is better revealed with weaker supervision strength. All experiments results are reproduced by us, except the blue star is the mean value reported in [16].*

4.1 Comparisons with State-Of-The-Art (SOTA)

Synthetic datasets: Our method achieves comparable results on MIC and TIC compared with the fully supervised learning methods CausalVAE [16] and ConditionVAE [15], and outperform other unsupervised learning methods. As shown in Table 1, comparing to CausalVAE and ConditionVAE, our method can achieve slightly better performance on PosMIC, PosTIC, NegMIC and NegTIC. Unsupervised methods achieve low value on NegMIC and NegTIC due to barely learning semantic information. Further, the result of using a few labels to train our method is included in Appendix.

Real datasets: Compared with synthetic datasets, there are 40 generative factors in CelebA dataset. If no label is available during training CelebA datasets, the search space for the model becomes intractable as there are 2^{40} different binary causal graphs for 40 factors. To decrease the difficulty, label information is needed to control the semantic factors which are encoded in each dimension of latent space [4]. For comprehensive comparison, our model and baselines are trained with $\{20\%, 40\%, 60\%, 80\%, 100\%\}$ of labeled data, and the remaining samples are unlabelled. As shown in Figures 6 and 7, our method consistently and significantly outperforms CausalVAE. Furthermore, with fewer labels, our method outperforms CausalVAE more appreciably.

5 Conclusion

In this work, we propose a novel architecture for causal representation learning with reduced supervision strength, exploiting the *do-operation* . We use a pair of images and apply *do-operation* to both latent cause and effect factors for new reconstructions. By comparing the new reconstructions after *do-operation* and the original inputs, the supervision strength is reduced. Furthermore, to better evaluate causal representation learning, we propose new metrics to address adequacy of existing metrics. We empirically demonstrate the advantages of our method on both synthetic and real datasets.

Acknowledgement: This material is based on research sponsored by Air Force Research Laboratory under agreement number FA8750-19-1-1000. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

[1] 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.

[2] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018.

[3] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders, 2019.

[4] Zunlei Feng, Xinchao Wang, Chenglong Ke, An-Xiang Zeng, Dacheng Tao, and Mingli Song. Dual swap disentangling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[5] 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 *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[6] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 26–28 Aug 2020.

[7] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

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

[9] Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.

[10] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019.

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

[12] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*, 2019.

[13] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[14] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[15] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[16] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9593–9602. Computer Vision Foundation / IEEE, 2021.

[17] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, 2018.

A Appendix

A.1 Importance of the *do-operation*

To prove the importance of the different *do-operation* modules used in our method, we evaluate our model by removing different *do-operation* modules in the architecture. As shown in Table A1, by removing Do-Cause, the model loses the ability of finding causal relationship. Removing Do-effect will lead to performance decrease on NegMIC and NegTIC.

Therefore, the performance on MIC, TIC, PosMIC and PosTIC degrades similar to unsupervised CausalVAE. By removing Do-effect and keeping Do-Cause, the performance on MIC, TIC, PosMIC and PosTIC significantly improves, while the performance on NegMIC and NegTIC is worse than full model where both cause and effect *do-operation* modules are used.

Table A1: *Causal representation metrics of model with different* do-operation *module applied*

Do-Cause	Do-Effect	MIC ↑	TIC ↑	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑
-	✓	30.6	25.9	23.6	17.2	**19.2**	**11.6**	36.5	28.8
✓	-	84.2	72.3	52.6	42.1	46.3	37.9	53.1	50.2
✓	✓	**86.6**	**74.5**	**54.1**	**44.0**	40.2	31.6	**56.8**	**53.6**

A.2 GAE comparison with NOTEARS

As we mentioned in Section 3, our method incorporate a graph autoencoder (GAE) [17] as causal discovery layer. GAE can learn nonlinear structural causal relationships thus generalizing over NOTEARS [29] which can only learn linear mapping. As shown in Table A2, if we replace GAE with NOTEARS for causal discovery layer, the performance of our model will be harmed since the causal relationships between latent factors can be nonlinear in many cases.

Table A2: *Causal representation metrics tested on Pendulum and Flow. Higher MIC, TIC, PosMIC and PosTIC value mean better performance. Lower NegMIC and NegTIC value mean better performance. Our methods are trained using only* 10% *of label.*

Models	Pendulum							
	MIC ↑	TIC ↑	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑
NOTEARS	40.3	30.9	27.3	17.3	**26.2**	**16.2**	39.6	28.7
Our method	**86.6**	**74.5**	**54.1**	**44.0**	40.2	31.6	**56.8**	**53.6**

A.3 Synthetic datasets experiments of using a few labels

In Section 4, we demonstrate our method can outperform other methods which do not use the label, and our method can achieve comparable performance compared with CausalVAE evaluated by PosMIC, PosTIC, NegMIC, and NegTIC. To test our method more comprehensively on synthetic datasets, we conduct experiments of our method, CausalVAE and ConditionVAE using only 10% of labels. As shown in Table A3, trained under only 10% of labeled data, CausalVAE and ConditionVAE are difficult to learn either good semantic meaning latent factors which is reflected by MIC and TIC, or attain true causal relationship between cause factors and effect factors, which is shown by PosMIC and PosTIC. As we discussed in Section 4, since CausalVAE fails to encode useful enough semantic factors information, it achieves a low value on NegMIC and NegTIC. ConditionVAE achieves low NegMIC and NegTIC because it aims at learning disentangled latent representation, where each latent factor is enforced to be independent of each other. Thus no causal relationship, correct or wrong, will be learned.

A.4 Experiments detail

The true causal graph of each datasets are shown in Figure A1.

We use one NVIDIA 1080 Ti GPU as our training and inference device. Following CausalVAE [26] architecture, we show the VAE architecture of synthetic datasets in Table A4 and VAE architecture

Table A3: *Causal representation metrics tested on Pendulum and Flow. Higher MIC, TIC, PosMIC and PosTIC value mean better performance. Lower NegMIC and NegTIC value mean better performance. Our methods are trained using only* 10% *of label.*

Models	Pendulum								Flow							
	MIC	TIC	PosMIC	PosTIC	NegMIC	NegTIC	F_1^{MIC}	F_1^{TIC}	MIC	TIC	PosMIC	PosTIC	NegMIC	NegTIC	F_1^{MIC}	F_1^{TIC}
All labels used																
CausalVAE [26]	**95.1**	**81.6**	53.0	43.4	46.6	37.0	53.2	51.4	72.1	56.4	45.1	36.7	43.3	33.7	47.3	33.6
ConditionVAE [22]	93.8	79.6	36.5	27.8	34.6	25.7	46.9	40.5	75.5	**56.5**	28.6	21.3	27.2	20.6	41.1	33.6
10% labels used																
CausalVAE [26]	64.7	55.9	39.4	30.7	37.6	28.2	48.3	43.0	53.2	46.7	30.6	22.5	30.3	21.7	42.5	35.0
ConditionVAE [22]	63.2	52.1	30.5	21.3	**29.4**	**24.6**	42.6	33.2	55.7	48.1	29.6	20.8	**26.7**	**20.1**	42.1	33.0
Our method	94.6	80.7	**70.2**	**59.5**	41.2	30.4	**63.9**	**63.9**	**75.7**	56.1	**60.3**	**51.8**	37.8	29.6	**61.2**	**59.7**

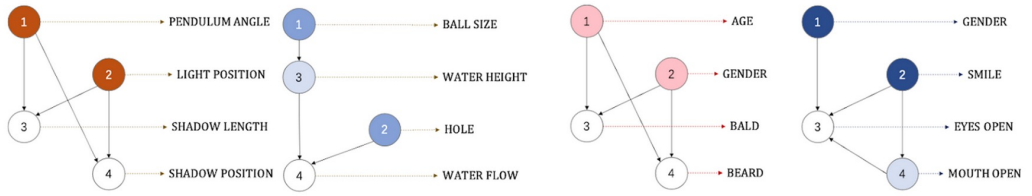


Figure A1: *Ground truth Causal graph of four datasets.*

of CelebA dataset in Table A5. For latent representation, we also follow the setting of CausalVAE where latent space z is extanded to matrix $z \in R^{n \times k}$ and n is the number of concept and k is latent dimension of each concept. k is set to 4 for VAE used in synthetic datasets and k is set to 32 for VAE used in CelebA dataset.

As described in Section 3, our loss function for no label training is shown in Equation 10 and the loss for label training is shown in Equation 11. The hyperparameters (α, β, γ) are grid search among $\{1e^{-3}, 1e^{-2}, 1e^{-1}, 1.0\}$. For training with label, the hyperparameter of l_u is always set to 1.

Table A4: *Synthetic datasets model architecture*

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

Table A5: *CelebA datasets model architecture*

encoder	decoder
-	1*1 conv. 128 1LReLU(0.2), stride 1
4*4 conv. 32 1LReLU (0.2), stride 2	4*4 convtranspose. 64 1LReLU(0.2), stride 1
4*4 conv. 64 1LReLU (0.2), stride 2	4*4 convtranspose. 64 1LReLU(0.2), stride 1
4*4 conv. 64 1LReLU (0.2), stride 2	4*4 convtranspose. 32 1LReLU(0.2), stride 1
4*4 conv. 64 1LReLU (0.2), stride 2	4*4 convtranspose. 32 1LReLU(0.2), stride 1
4*4 conv. 256 1LReLU (0.2), stride 2	4*4 convtranspose. 32 1LReLU(0.2), stride 1
1*1 conv. 3, stride1	4*4 convtranpose. 3, stride 2

A.5 Do-operation implementation detail

As we described in section 3, we apply *do-operation* to both latent cause and effect factors. To better show the implementation of *do-operation* in our work, we describe the process in Figure A2. As illustrated in Figure A2, the cause and effect factors in the latent space are decided by learned causal matrix A which is identical to causal matrix used in causal discovery layer. After deciding the cause and effect factors, we separately apply *do-operation* on cause and effect factors. Applying *do-operation* to cause factors is straightforward since cause factors have no parent factors and the

causal graph stays unchanged. Oppositely, applying *do-operation* to effect factors will both fix the value of effect factors and remove affects from cause factors. Thus, if we swap the effect factors before causal discovery layer, the original causal relationships from cause factors to effect factors still hold. To eliminate the original causal relationships, the swapping operation on effect factors should be applied after causal discovery layer.

According to [18], *do-operation* replace factors with constants and remove all causal relationships towards the factors. If the label information is available, the *do-operation* is straightforward since the latent factors value can be easily fixed with the label value. However, if the label information is missing, even though the latent factor value can be replaced by some random values, such random values do not guarantee to be meaningful. To obtain the proper constants which replace latent factors, another sample is needed since the reconstruction task force the latent representation encoded from the input are meaningful and can be used as source for *do-operation*.

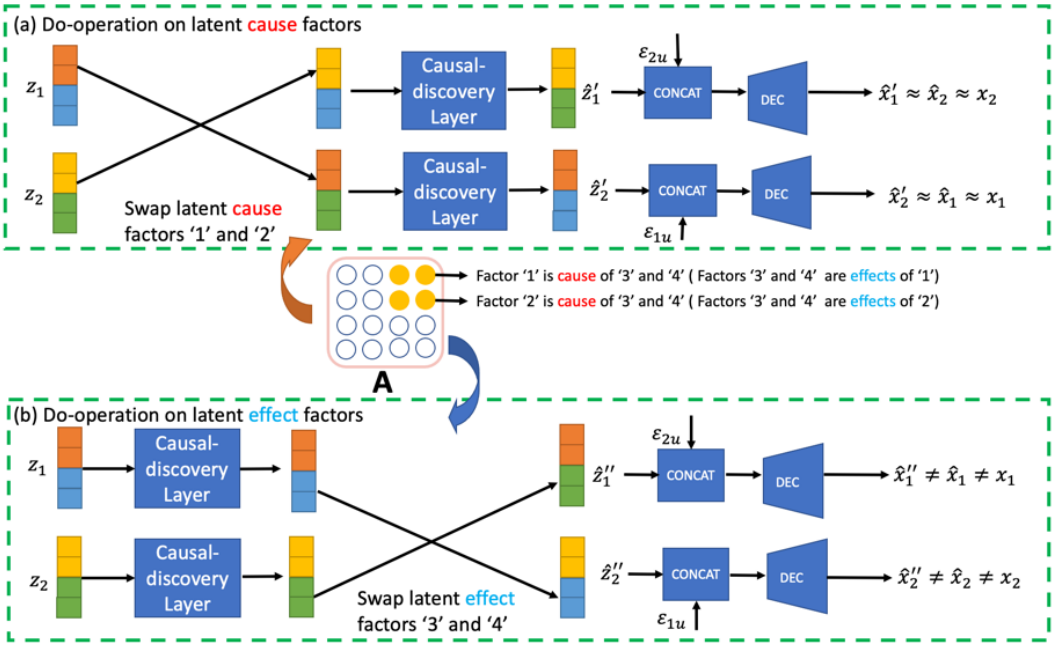


Figure A2: Do-operation *is applied to both cause factors and effect factors.* Do-operation *on cause factors encourage model to learn correct causal relationships and* do-operation *on effect factors prevent model learning wrong causal relationships.*

A.6 Counter example to prove the weakness of MIC and TIC

Assuming we have four independent gaussian variables A, B, C and D, where $A \sim \mathcal{N}(\mu_a, \sigma_a^2)$, $B \sim \mathcal{N}(\mu_b, \sigma_b^2)$, $C \sim \mathcal{N}(\mu_c, \sigma_c^2)$ and $D \sim \mathcal{N}(\mu_d, \sigma_d^2)$. We can create other four gaussian variables A', B', C' and D' where $A' \sim \mathcal{N}(\mu_a, \sigma_a^2)$, $B' = \frac{\mu_b}{\mu_a} \cdot A' + (\sigma_b - \frac{\mu_b}{\mu_a}\sigma_a) \cdot \mathcal{N}(0,1)$, $C' = \frac{\mu_c}{\mu_a} \cdot A' + (\sigma_c - \frac{\mu_c}{\mu_a}\sigma_a) \cdot \mathcal{N}(0,1)$ and $D' = \frac{\mu_d}{\mu_a} \cdot A' + (\sigma_d - \frac{\mu_d}{\mu_a}\sigma_a) \cdot \mathcal{N}(0,1)$. By creating new variables like this, it is easy to see that A' has same distribution with A, B' has same distribution with B, C' has the same distribution with C and D' has the same distribution with D. Since MIC and TIC only evaluate the marginal distribution of each variable separately, they can not distinguish A from A', B from B', C from C' and D from D'. However, (A, B, C, D) have totally different joint distribution from (A', B', C', D').

A.7 Metrics implement details

The superiorities of the proposed new metrics and a simple example has been discussed in Section 4. More details about those new metrics will be discussed in this section. For fully supervised learning or semi-supervised learning method, the metrics calculation are straightforward since every latent elements is controlled by their corresponding label information [5]. For unsupervised methods and

our reduced supervision method without using label, we have to first determine the correspondence between each latent factor and each label. We use MIC to choose which latent element represent the label information. As we described in Section 4, MIC can be used to measure the information relevance between a latent factor and a generative label. For each generative factor label, we choose the latent element which achieve maximum MIC value evaluated with that generative factor. After choosing the correspondence between each latent factor with all generative factors label, we can apply Pos/Neg metrics according to the true causal graph provided by the datasets.

A.8 Reconstruction results

We include the image reconstruction results in this section. Shown in figs. A3 to A6, when changing the cause factors, the effect factors shown in reconstructions are changed corresponding. On the contrary, when changing the effect factors, the reconstructions can be counterfactual images and the cause factors stay unchanged.

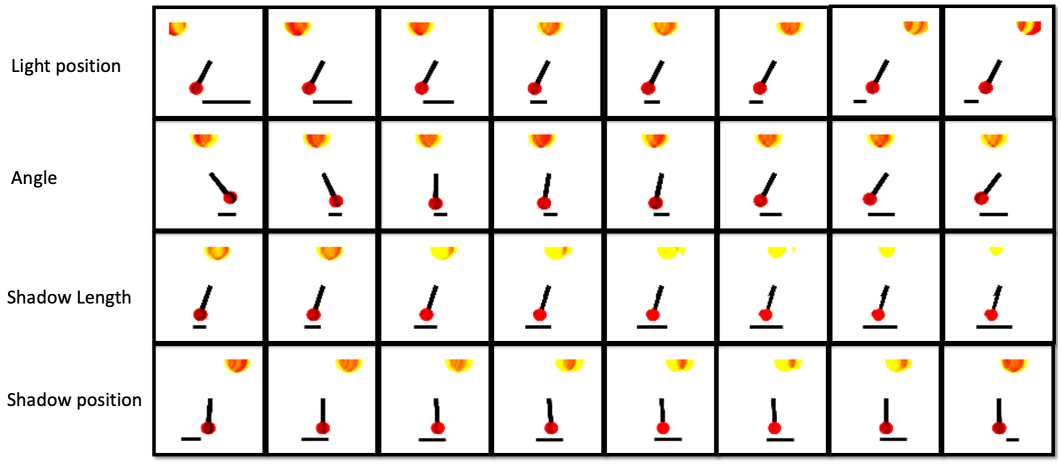


Figure A3: *Traversal reconstruction of pendulum dataset. For each rows, we only change one latent factor value and fix all other latent factors. By changing cause factor (light position or angle), we observe corresponding change in effect factors (shadow position and shadow length). Oppositely, by changing effect factor (shadow location and shadow length), the reconstructions can become counterfactual images and the cause factors (light position and angle) stay unchanged.*

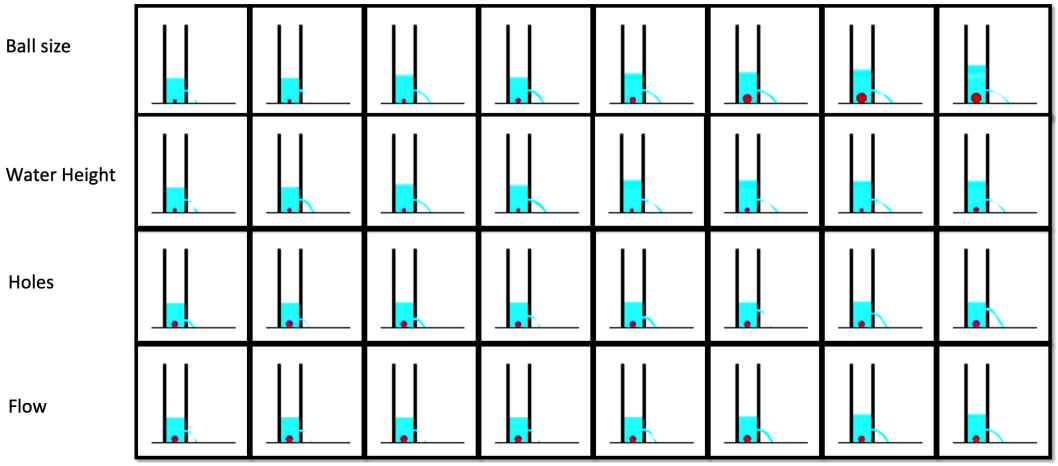


Figure A4: *Traversal reconstruction of flow dataset. For each rows, we only change one latent factor value and fix all other latent factors. By changing cause factor (ball size or hole), we observe corresponding change in effect factors (water height and flow). Oppositely, by changing effect factor (water height or flow), the reconstructions can become counterfactual images and the cause factors (ball size and hole) stay unchanged.*

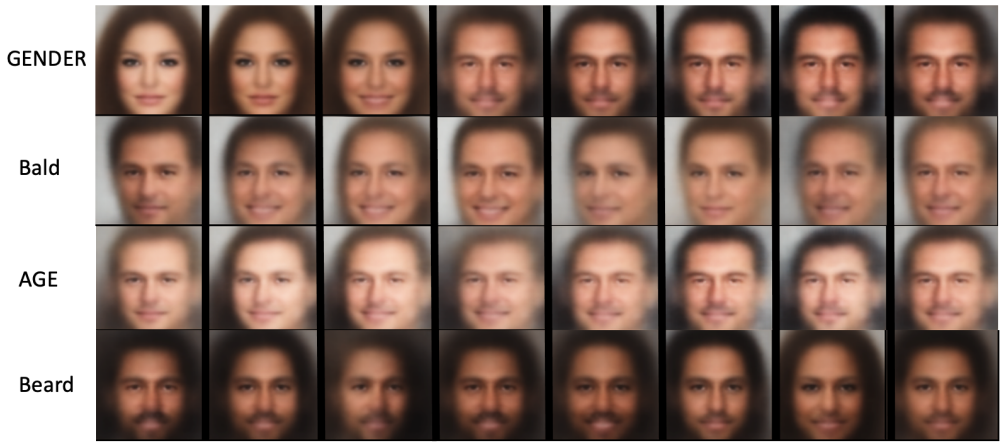


Figure A5: *Traversal reconstruction of CelebA(Beard) dataset. For each rows, we only change one latent factor value and fix all other latent factors. By changing cause factor (age or gender), we observe corresponding change in effect factors (bald and beard). Oppositely, by changing effect factor (beard and bald), the reconstructions can become counterfactual images and the cause factors stay unchanged.*

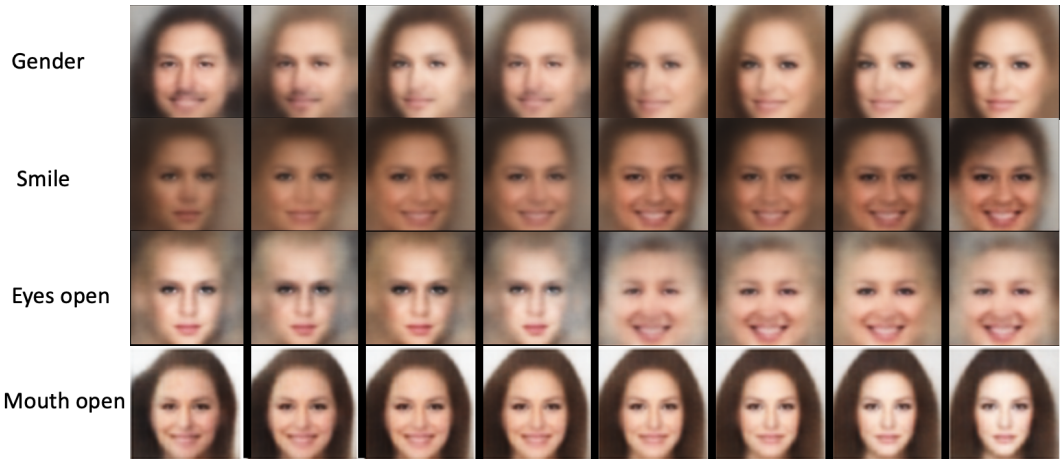


Figure A6: *Traversal reconstruction of pendulum CelebA(Smile) dataset. For each rows, we only change one latent factor value and fix all other latent factors. By changing cause factor (gender and smile), we observe corresponding change in effect factors (eyes open). Oppositely, by changing effect factor (shadow eyes open), the reconstructions can become counterfactual images and the cause factors stay unchanged.*