

# Causally Disentangled Generative Variational AutoEncoder

26th European Conference on Artificial Intelligence ECAI 2023

---

Seunghwan An<sup>1</sup>, Kyungwoo Song<sup>2</sup>, and Jong-June Jeon<sup>1\*</sup>

<sup>1</sup>Department of Statistics, University of Seoul, S. Korea

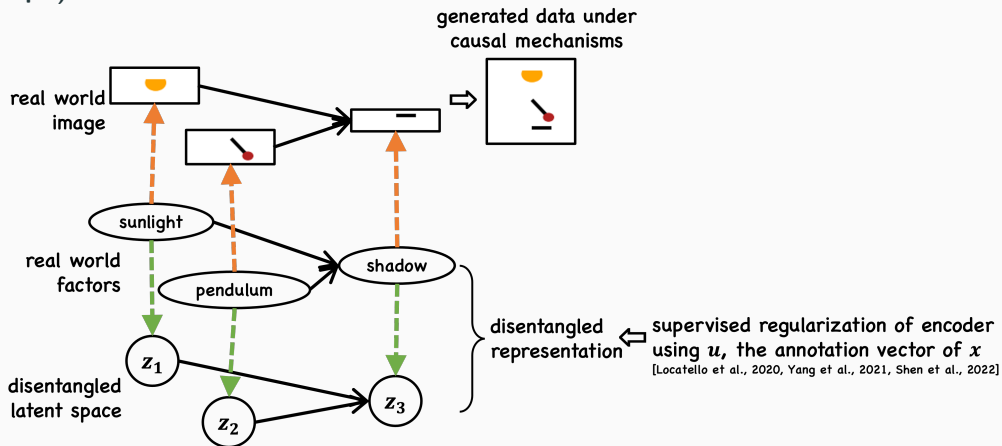
<sup>2</sup>Department of Applied Statistics, Department of Statistics and Data Science, Yonsei University, S. Korea

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

## Introduction

---

## (Example) Pendulum dataset <sup>1</sup>



Goal: **Learning a causally disentangled (causally-aware) generative model**

## Contribution:

The disentangled decoder is required to achieve the causally disentangled generative model.

- Assumption (in this presentation): The disentangled representation is already obtained.
  - $\mathbf{z}_1$ : sunlight,  $\mathbf{z}_2$ : pendulum,  $\mathbf{z}_3$ : shadow

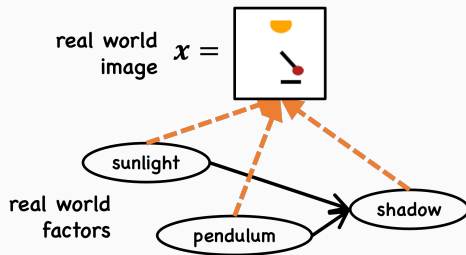
## Proposal: CDG-VAE

---

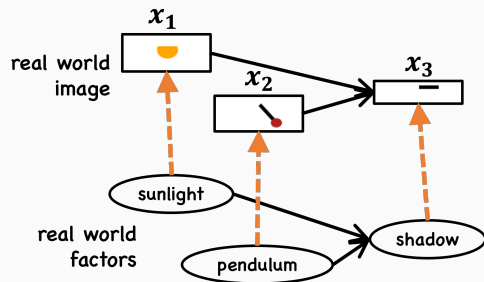
## Assumption 1 (Blocked representation)

There are (block) causal relationships,  $g_j \rightarrow x_j$ , where  $g$  is the ground-truth factors and  $j = 1, \dots, d$ .

### Conventional representation



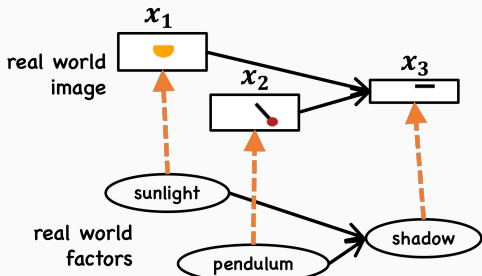
### Blocked representation



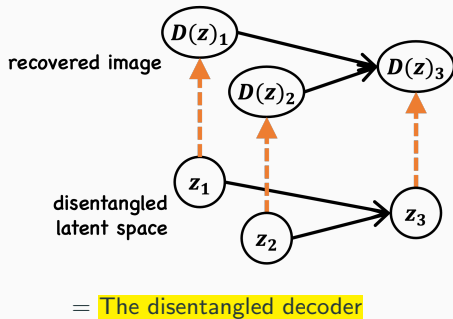
## Definition 1 (Causally Disentangled Generation (CDG))

Let  $D(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^p$  be a decoder mean vector of the model where the input is denoted as  $\mathbf{z} \in \mathbb{R}^d$ . Then the model is causally disentangled generative if, for  $i = 1, \dots, d$ ,  $D(\mathbf{z})_i$  is independent to  $\mathbf{z}_s, s \neq i$ , given  $\mathbf{z}_i$ .

### Ground-truth DGP

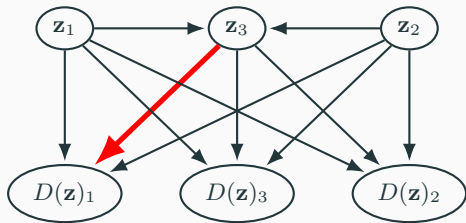


### DGP of the decoder

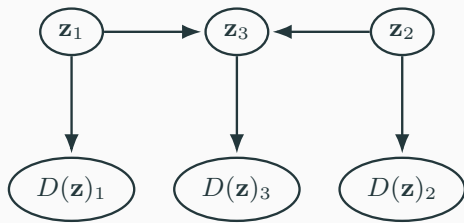


# Why do we need the disentangled decoder?

Entangled decoder  $\Rightarrow$  Causally Implausible



Disentangled decoder  $\Rightarrow$  Causally Plausible



## Definition 2 (Average Causal Effect (ACE))

Suppose that  $z_i, i = 1, \dots, d$  is intervened with  $z_i^{(1)}$  and  $z_i^{(2)}$ . Then, for  $c = 1, \dots, d$ , the average causal effect of  $z_i$  on the annotation  $\mathbf{u}_c$  given  $z_{ND(i)}$  is defined as

$$ACE(\mathbf{u}_c, z_i, z_{ND(i)} = z_{ND(i)}) := \left| \mathbb{E}[\mathbf{u}_c | z_{ND(i)}, do(z_i := z_i^{(1)})] - \mathbb{E}[\mathbf{u}_c | z_{ND(i)}, do(z_i := z_i^{(2)})] \right|.$$

$\Rightarrow$  The causal effect is measured by the difference of the annotation vector.



## How can we construct the disentangled decoder?

### Proposition 1 (Sufficient Condition for CDG)

Let  $D(\cdot; \theta) : \mathbb{R}^d \mapsto \mathbb{R}^p$  be a decoder mean vector of the model. If the decoder structure of the model satisfies  $D(\mathbf{z}; \theta) := (D(\mathbf{z}_1; \theta_1), \dots, D(\mathbf{z}_d; \theta_d))$ , where  $\theta = (\theta_1, \dots, \theta_d)$ , and  $D(\cdot; \theta_j)$  is a function parameterized with  $\theta_j$  for  $j = 1, \dots, d$ , then the model satisfies Definition 1.

- a generated image

$$\begin{aligned} D(\mathbf{z}; \theta) &= \left( D(\mathbf{z}; \theta)_1, \quad D(\mathbf{z}; \theta)_2, \quad D(\mathbf{z}; \theta)_3 \right) \\ &\Downarrow \\ D(\mathbf{z}; \theta) &= \left( \underbrace{D(\mathbf{z}_1; \theta_1)}_{\text{sunlight image}}, \quad \underbrace{D(\mathbf{z}_2; \theta_2)}_{\text{pendulum image}}, \quad \underbrace{D(\mathbf{z}_3; \theta_3)}_{\text{shadow image}} \right) \end{aligned}$$

# What is the property of CDG?

## Proposition 2 (Necessary Conditions for CDG)

For  $i = 1, \dots, d$ , assume that arbitrary  $x$  and  $z_{ND(i)}$  are given.  $z_{(i, z_{ND(i)}, x)}^{(j)}$  denotes the value of  $\mathbf{z}$  under intervention  $do(\mathbf{z}_i := z_i^{(j)})$  given  $x$  and  $z_{ND(i)}$ , for  $j = 1, 2$ . For  $c = 1, \dots, d$ , under the faithfulness assumption, if the model satisfies CDG (Definition 1) and

1.  $c \in ND(i)$ , then

$$ACE(\mathbf{u}_c, \mathbf{z}_i, z_{ND(i)}) = 0.$$

$\Rightarrow$  Non-descendants are NOT affected by the intervention.

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

$$0 < ACE(\mathbf{u}_c, \mathbf{z}_i, z_{ND(i)}) \leq \mathbb{E}_{p(\mathbf{x})} \left| \mathbb{E}[\mathbf{u}_c | z_{(i, z_{ND(i)}, \mathbf{x})}^{(1)}] - \mathbb{E}[\mathbf{u}_c | z_{(i, z_{ND(i)}, \mathbf{x})}^{(2)}] \right|,$$

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

$\Rightarrow$  Descendants are affected by the intervention.

### Definition 3 (Causal Disentanglement Metric (CDM))

For  $c, i = 1, \dots, d$ , the causal disentanglement metric (CDM) is defined as

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

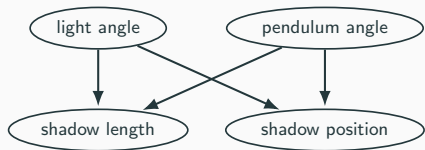
where  $\mathbb{E}$  indicates the expectation with respect to  $\mathbf{z}_{ND(i)}$ .

⇒ Expected value of the average causal effect.

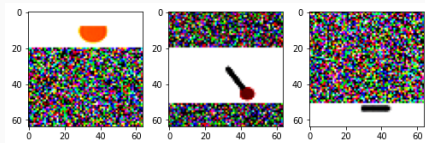
1. **interventional robustness** [Suter et al., 2019]
2. **counterfactual generativeness** [Reddy et al., 2022]

## Experiments

---

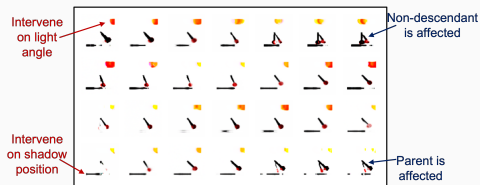


(a)



(b)

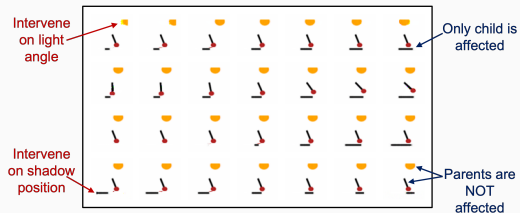
**Figure 1:** (a) DAG of the ground-truth factors:  $g_1$ (light angle),  $g_2$ (pendulum angle),  $g_3$ (shadow length), and  $g_4$ (shadow position). (b) From left to right,  $x_1$  (light),  $x_2$  (pendulum), and  $x_3$  (shadow).



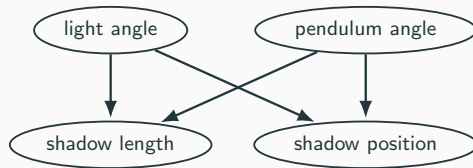
(a) CausalVAE [Yang et al., 2021]



(b) DEAR [Shen et al., 2022]



(c) CDG-VAE



(d) DAG of the ground-truth factors

⇒ CDG-VAE under Proposition 1 enables causally disentangled generation!





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


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

## References

---



-  Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. (2020).  
**Disentangling factors of variations using few labels.**  
In *International Conference on Learning Representations*.
-  Reddy, A. G., Balasubramanian, V. N., et al. (2022).  
**On causally disentangled representations.**  
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8089–8097.
-  Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. (2022).  
**Weakly supervised disentangled generative causal representation learning.**  
*Journal of Machine Learning Research*, 23:1–55.
-  Suter, R., Miladinovic, D., Schölkopf, B., and Bauer, S. (2019).  
**Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness.**  
In *International Conference on Machine Learning*, pages 6056–6065. PMLR.

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

Thank you!



**Figure 2:** GitHub repository link of CDG-VAE.