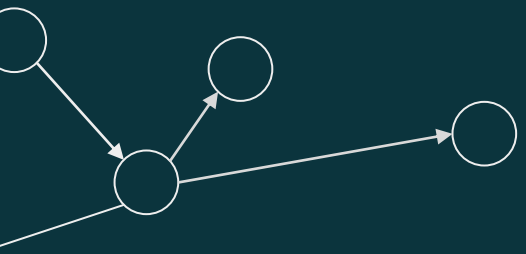
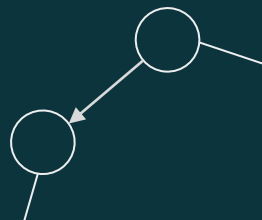
waiting for the lecture to start

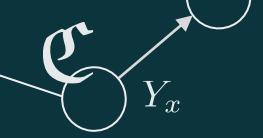


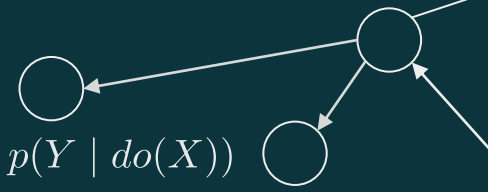
TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

Causal Representation Learning

Devendra Singh Dhami

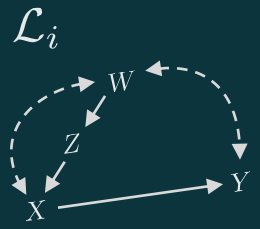






Section

I Motivation



Motivation

You have learned about causal discovery.
Given such an image, what is the causal graph?

Motivation

What about this image?

Motivation

Or this?

Motivation

We can group parts of the image that “belong together”
(here: Segment Anything [1])

Use these as variables?

[1] Kirillov, Alexander, et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

Motivation

We can group parts of the image that “belong together”
(here: Segment Anything [1])

Use these as variables?

But what should a variable be?
→ More than just “pixel labeling”

[1] Kirillov, Alexander, et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

Motivation

Another example: A classifier (clf) tries to distinguish between cows and camels

Training Data

Camels



Cows



Clf

Motivation

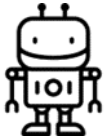
Another example: A classifier tries to distinguish between cows and camels

Test Data

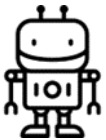
Motivation

Another example: A classifier tries to distinguish between cows and camels

Test Data



Clearly a cow!



Motivation

Another example: A classifier tries to distinguish between cows and camels

Test Data

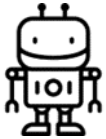
Clearly a cow!

A camel, obviously!

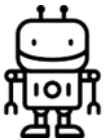
Motivation

Another example: A classifier tries to distinguish between cows and camels

Test Data



Clearly a cow!

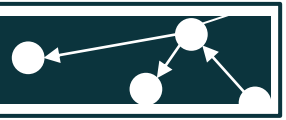


A camel, obviously!

- Use disentangled representations for OOD classification
- Relations are important:

Camels usually live in a desert but a camel on grass is still a camel.

Motivation

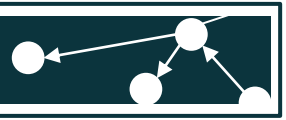


CVs: Causal Variables
CG: Causal Graph

We have seen:

- Causal Inference — known CVs ✅ , CG ✅
- Causal Structure Learning — known CVs ✅ , CG ❌

Motivation



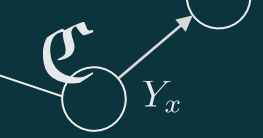
CVs: Causal Variables
CG: Causal Graph

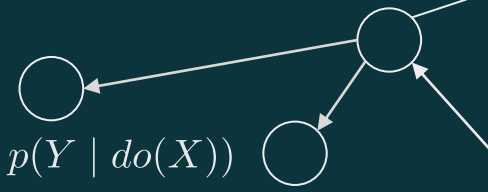
We have seen:

- Causal Inference — known CVs ✅ , CG ✅
- Causal Structure Learning — known CVs ✅ , CG ❌

Today:

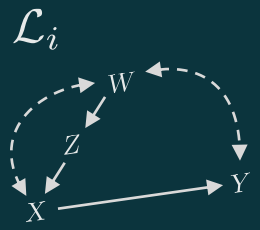
- **Causal Representation Learning — known CVs ❌, CG ❌**





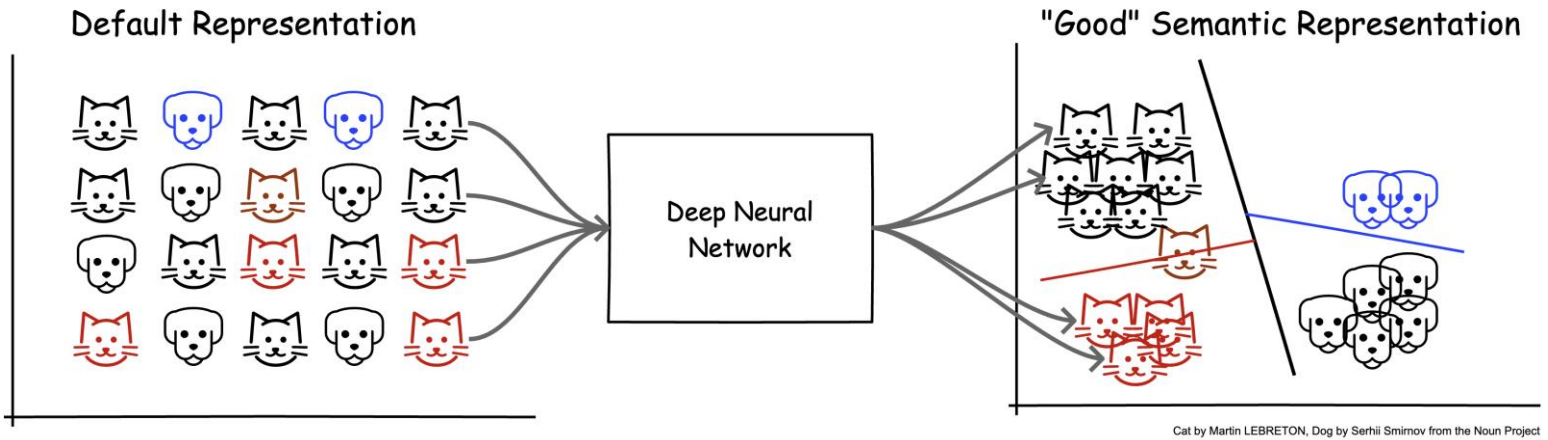
Section

2 Representation Learning Basics



Representation Learning

Representation Learning: “**learning representations** of the data that make it easier to extract useful information when building classifiers or other predictors” [2]

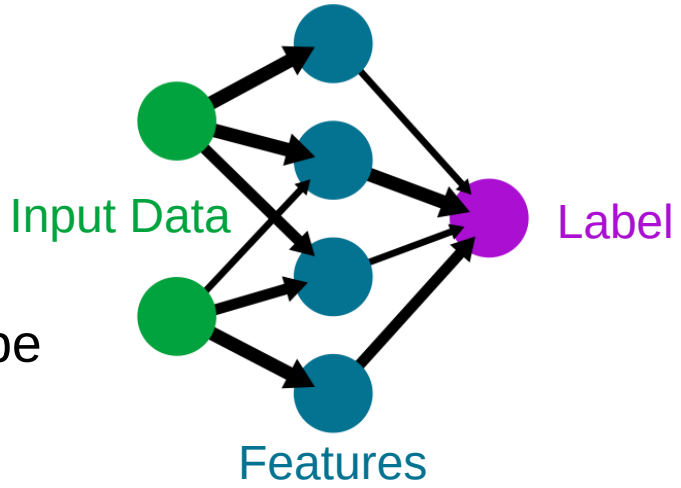


[2] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

Representation Learning

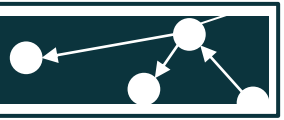
Representation Learning: "**learning representations** of the data that make it easier to extract useful information when building classifiers or other predictors" [2]

- Also: Feature Learning
- Technically most if not all of deep learning incorporates feature learning
- Not exclusive to images (though these will be the focus of this lecture)

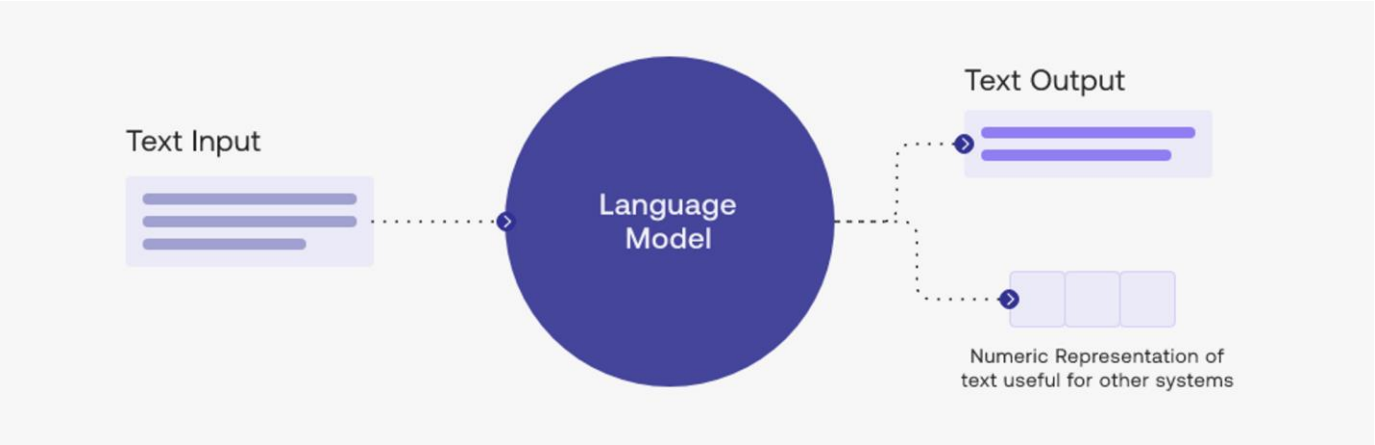


[2] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

Representation Learning



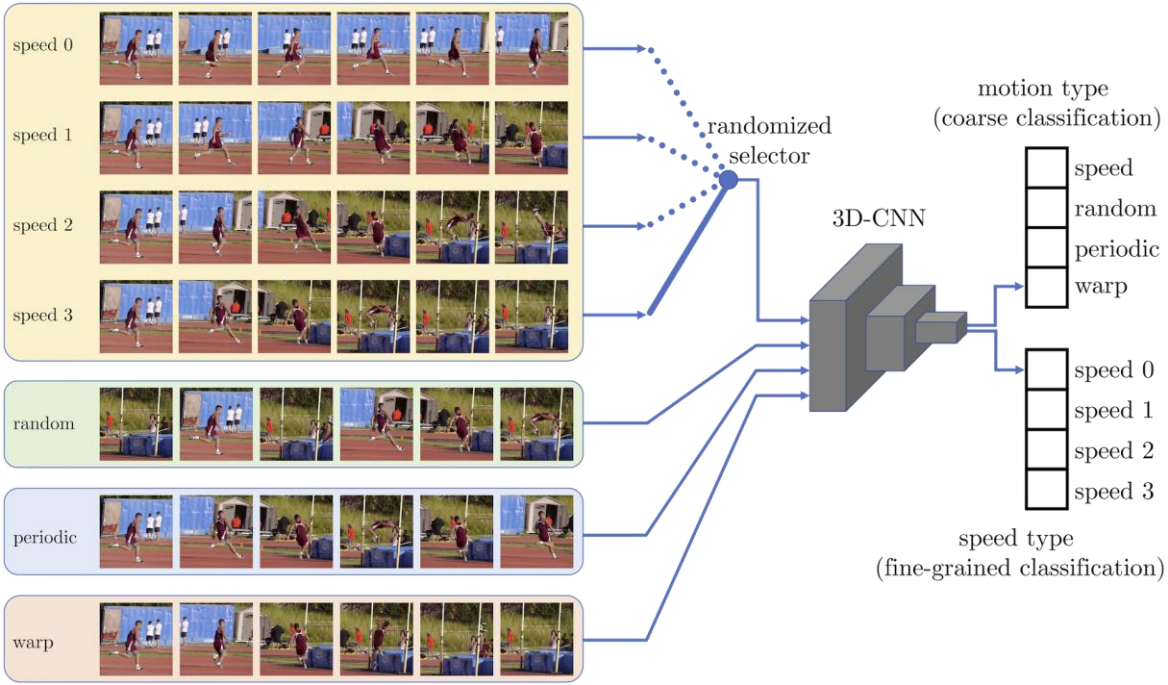
- Not exclusive to images (though these will be the focus of this lecture)



https://attri.ai/blog/introduction-to-large-language-models

Representation Learning

- Not exclusive to images (though these will be the focus of this lecture)



Jenni et al., Video Representation Learning by Recognizing Temporal Transformations, ECCV 2020

Representation Learning

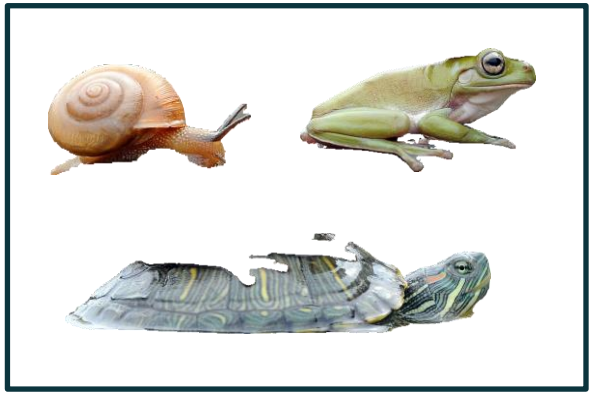
- Data $X \in \mathbb{R}^p$ can be described a set of latent variables $Z \in \mathbb{R}^d$ where $d \ll p$. Then, $Z = h(X)$ and $X = g(Z)$.

Illustrative example:

$X \in \mathbb{R}^{2500x1661}$

$h(X)$

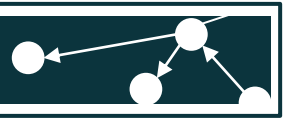
$g(Z)$



$Z \in \mathbb{R}^{1000}$ (*)

(*) let's say there are 1000 classes that can be either true or false

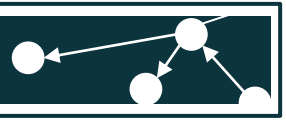
Representation Learning



- Data $X \in \mathbb{R}^p$ can be described a set of latent variables $Z \in \mathbb{R}^d$ where $d \ll p$. Then, $Z = h(X)$ and $X = g(Z)$.

$$h \stackrel{?}{=} g^{-1}$$

Representation Learning

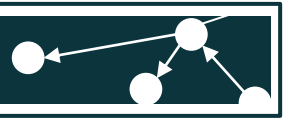


- Data $X \in \mathbb{R}^p$ can be described a set of latent variables $Z \in \mathbb{R}^d$ where $d \ll p$. Then, $Z = h(X)$ and $X = g(Z)$.

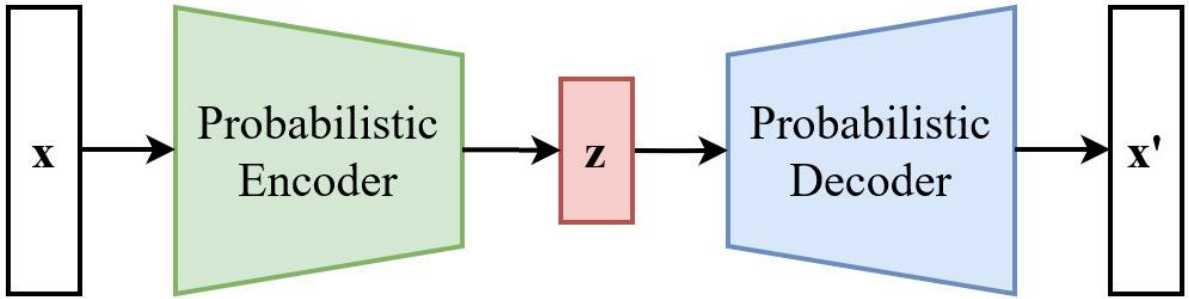
$$h \stackrel{?}{=} g^{-1}$$

- If g is invertible, $Z = g^{-1}(X)$ is a perfect representation of X
- However, due to $d \ll p$, g can not be invertible if we want to capture the full space of $X \in \mathbb{R}^p$

Representation Learning

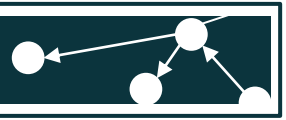


Autoencoder



- Two neural networks: encoder $X \rightarrow Z$ and decoder $Z \rightarrow X$

Representation Learning



What does an autoencoder achieve? What can it not be used for?

Representation Learning

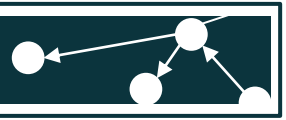
What does an autoencoder achieve? What can it not be used for?

✅ Learn compact encoded representation that can be used to reconstruct the original input with minimal error

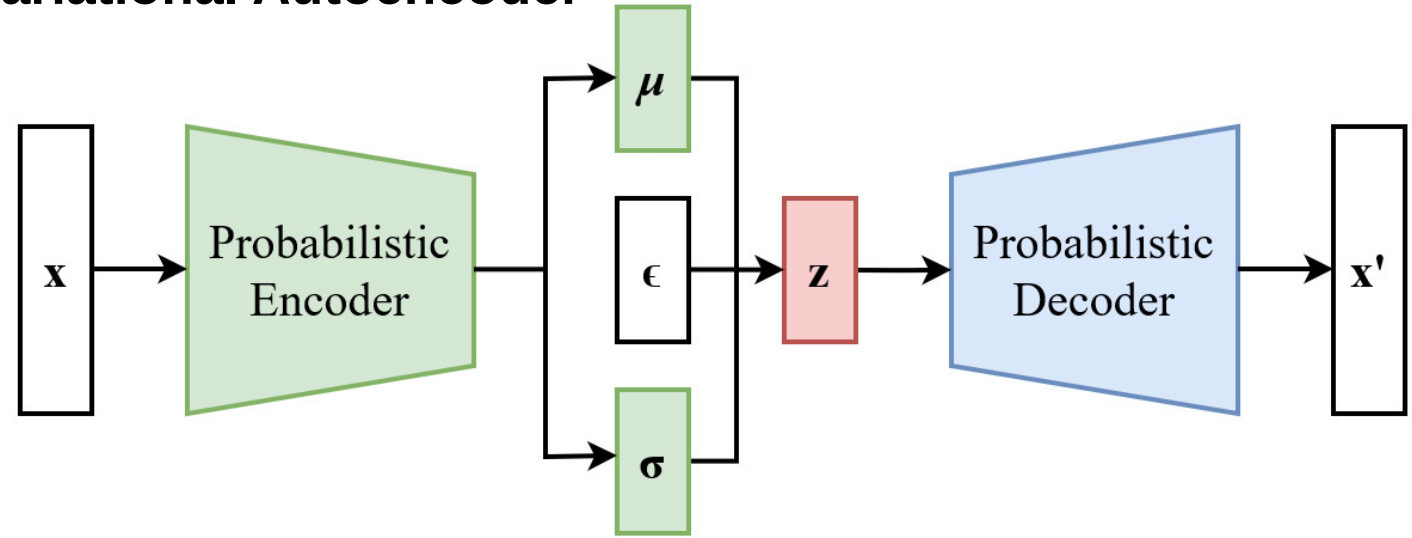
❌ Encoded representation is not guaranteed to have any tangible meaning; no disentanglement

❌ It is difficult if not impossible to interpret or purposefully “intervene” on the encoded representation

Representation Learning

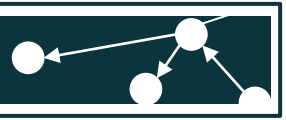


Variational Autoencoder



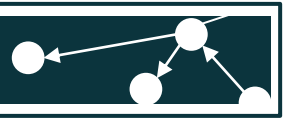
- The latent is encoded as a distribution instead of a single vector
- Trained with loss for reconstruction and regularization

Representation Learning

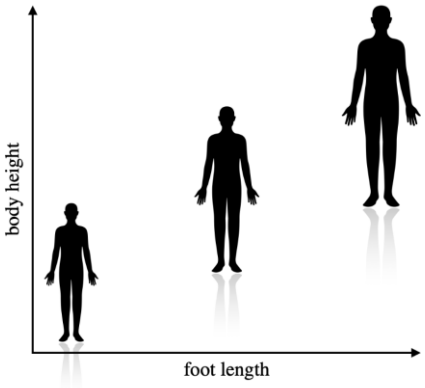


There are many good resources going into more depth on the idea and mathematical details of variational autoencoders which were skipped due to time constraints. We recommend looking into these in case of interest.

Representation Learning



- Autoencoders work well for uncorrelated features
- What if features are correlated?
- Example (taken from [3]): autoencoder learns two latent features

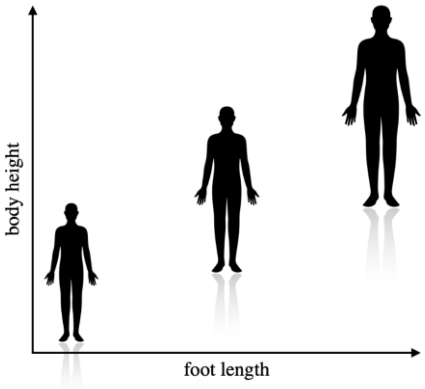


Correlated Features:
Body Height and Foot Length

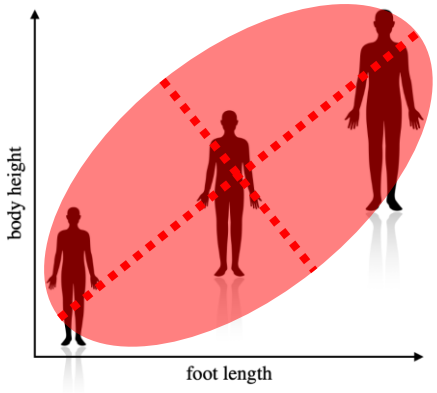
[3] Träuble, Frederik, et al. "On disentangled representations learned from correlated data." *International Conference on Machine Learning*, 2021.

Representation Learning

- Autoencoders work well for uncorrelated features
- What if features are correlated?
- Example (taken from [3]): autoencoder learns two latent features

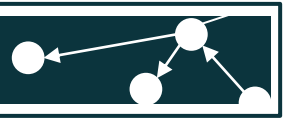


Correlated Features:
Body Height and Foot Length

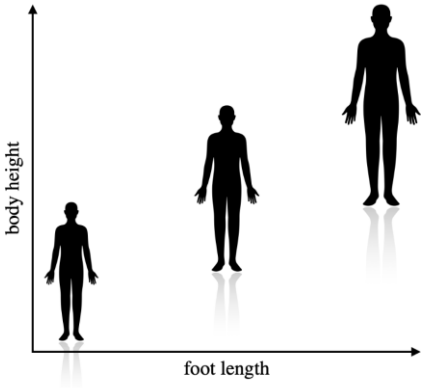


Entangled
Features

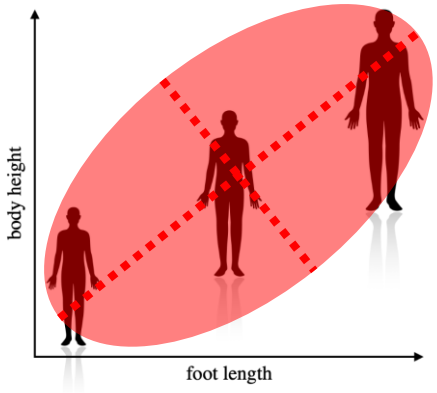
Representation Learning



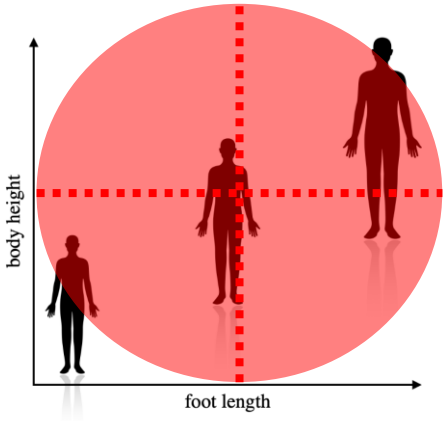
- Autoencoders work well for uncorrelated features
- What if features are correlated?
- Example: autoencoder learns two latent features



Correlated Features:
Body Height and Foot Length

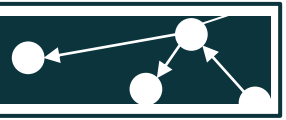


Entangled
Features



Disentangled
Features

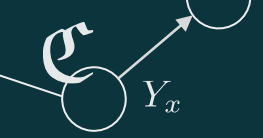
Representation Learning

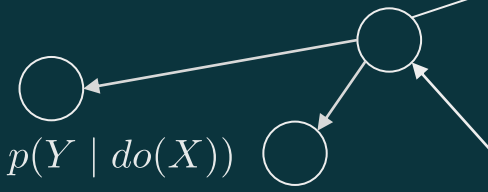


- Disentangled features do not match true probability density as well as entangled representation
- Probability mass is placed outside of true (train) distribution
- Independency assumption of latent features does not hold

- We want to learn interpretable, disentangled features representations that take the structure of the real world into account

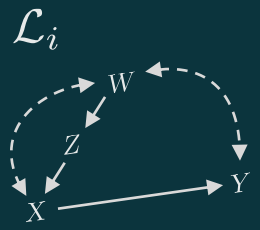
→ Therefore…



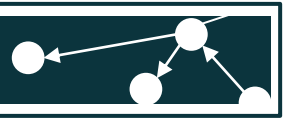


Section

3 Causal Representation Learning (CRL) Basics



CRL Basics



What is Causal Representation Learning?

CVs: Causal Variables
CG: Causal Graph

- Causal Inference — known CVs ✅, CG ✅
- Causal Structure Learning — known CVs ✅, CG ❌
- **Causal Representation Learning** — **known CVs ❌, CG ❌**

CRL Basics

What is Causal Representation Learning? (figure taken from [4])

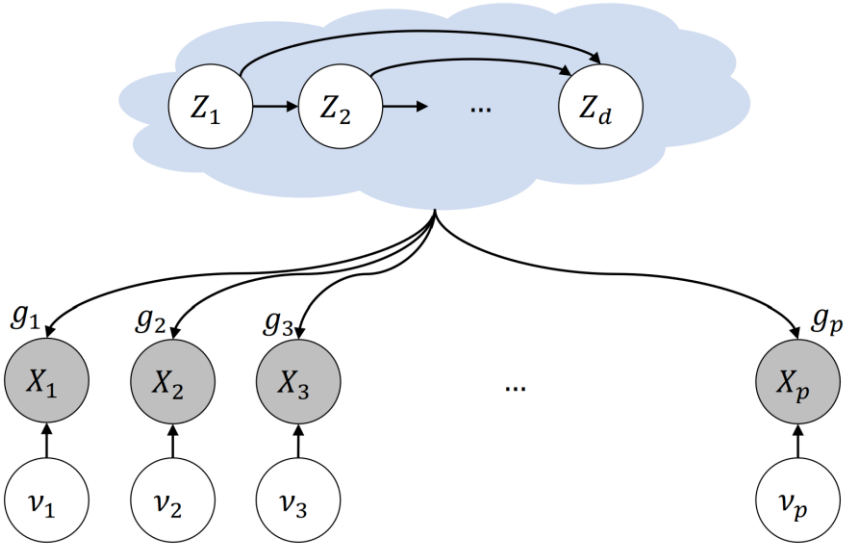


Figure 7.1: The causal disentanglement model.

[4] Squires, Chandler. “Chapter 7 Causal Representation Learning” 6.S091-causality, 2023, https://github.com/csquires/6.S091-causality/blob/main/lecture_notes/Lecture7.pdf

CRL Basics

What is Causal Representation Learning? (figure taken from [4])

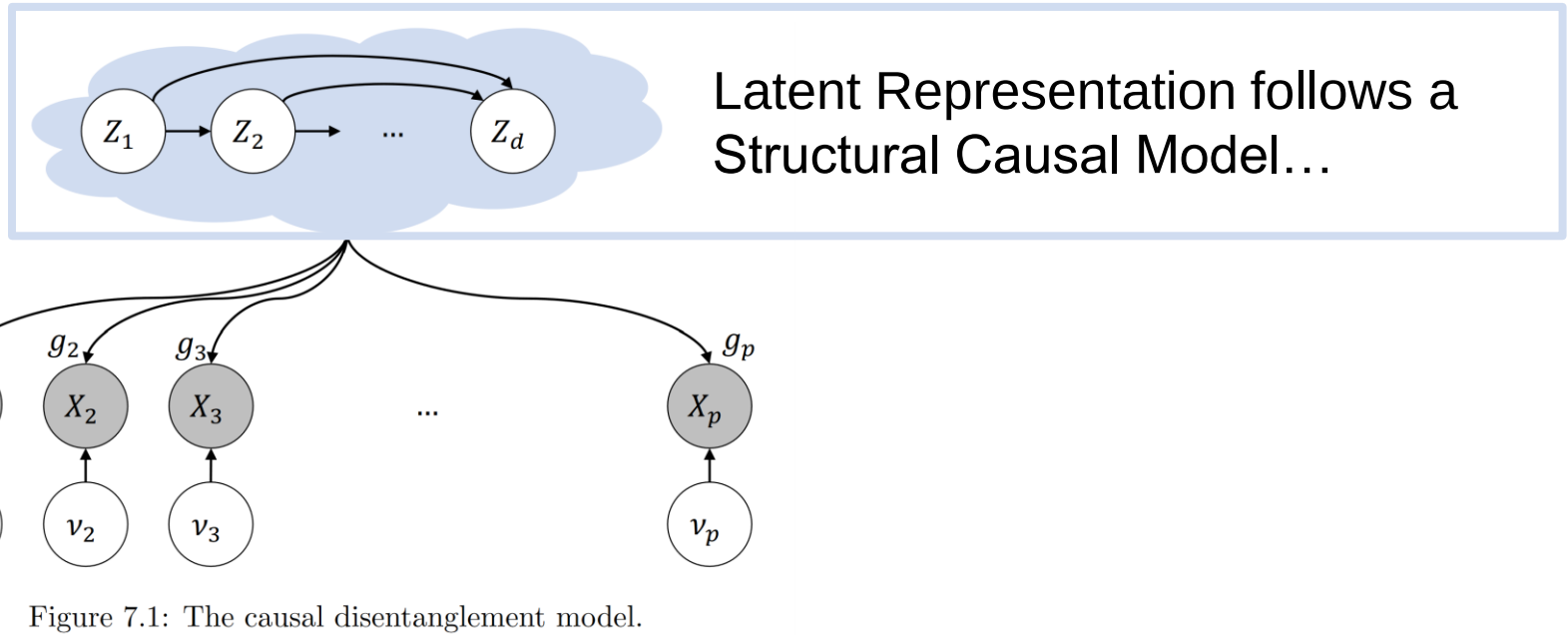


Figure 7.1: The causal disentanglement model.

[4] Squires, Chandler. “Chapter 7 Causal Representation Learning” 6.S091-causality, 2023, https://github.com/csquires/6.S091-causality/blob/main/lecture_notes/Lecture7.pdf

CRL Basics

What is Causal Representation Learning? (figure taken from [4])

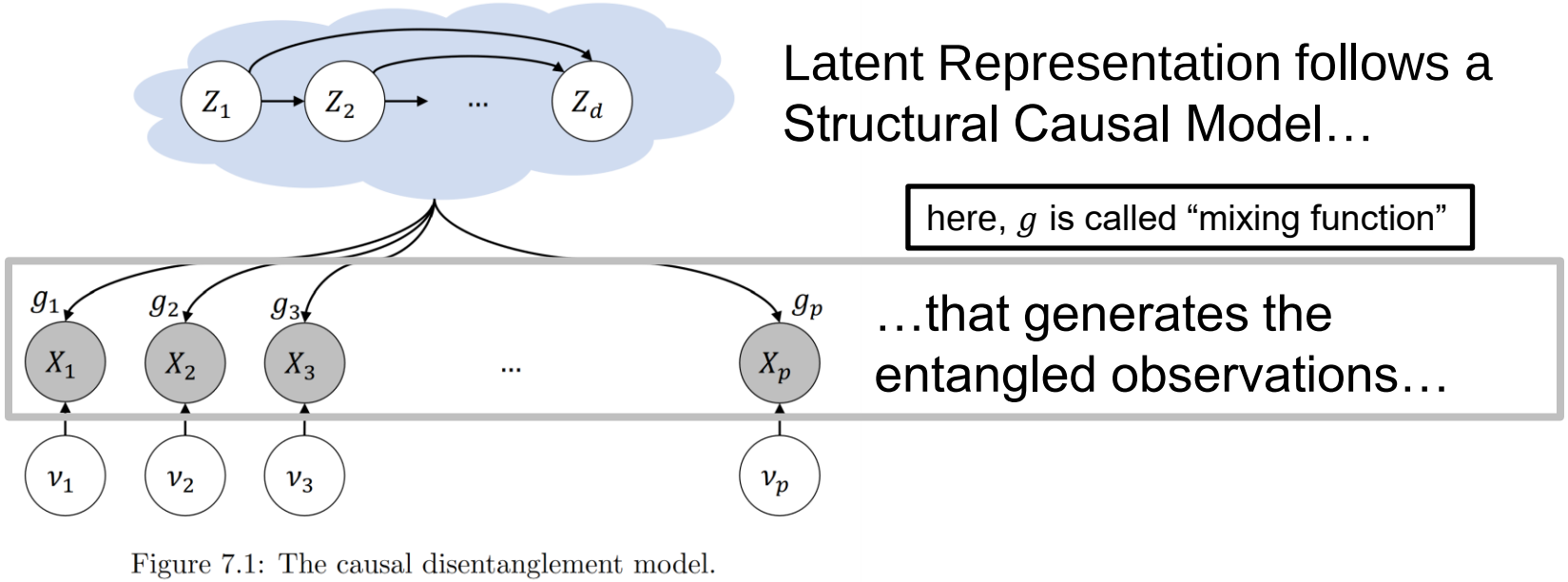


Figure 7.1: The causal disentanglement model.

Latent Representation follows a Structural Causal Model…

here, g is called “mixing function”

…that generates the entangled observations…

[4] Squires, Chandler. “Chapter 7 Causal Representation Learning” 6.S091-causality, 2023, https://github.com/csquires/6.S091-causality/blob/main/lecture_notes/Lecture7.pdf

CRL Basics

What is Causal Representation Learning? (figure taken from [4])

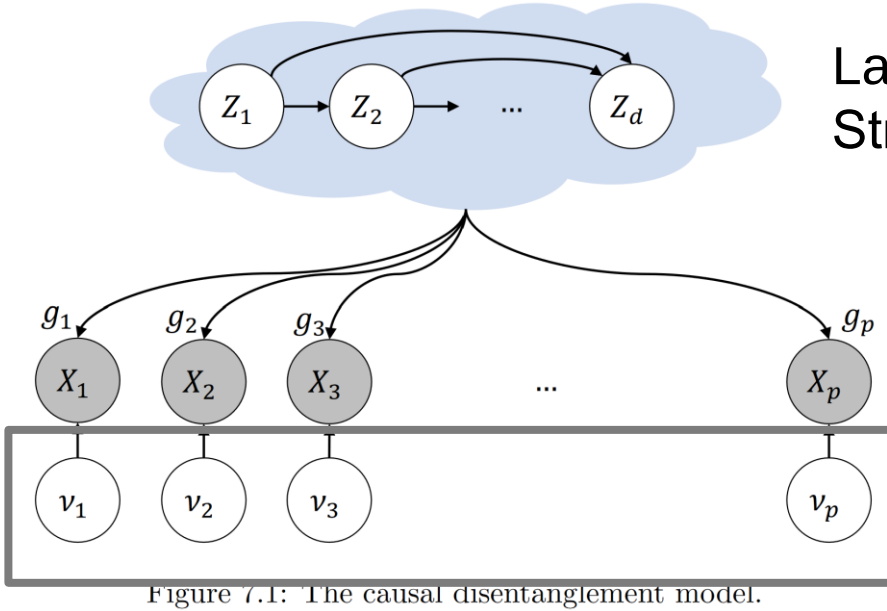


Figure 7.1: The causal disentanglement model.

Latent Representation follows a Structural Causal Model…

here, g is called “mixing function”

…that generates the entangled observations…

…which are also influenced by random noise

[4] Squires, Chandler. “Chapter 7 Causal Representation Learning” 6.S091-causality, 2023, https://github.com/csquires/6.S091-causality/blob/main/lecture_notes/Lecture7.pdf

CRL Basics

What is Causal Representation Learning? (figure taken from [4])

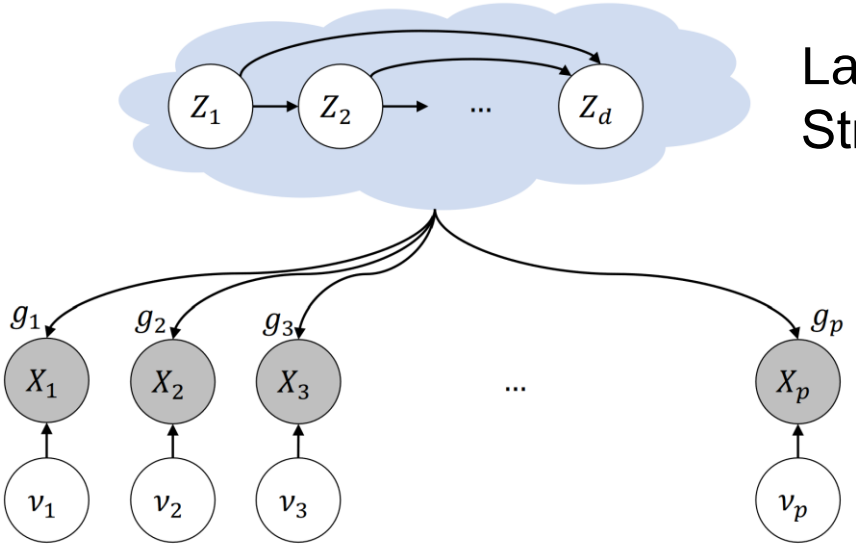


Figure 7.1: The causal disentanglement model.

Latent Representation follows a Structural Causal Model…

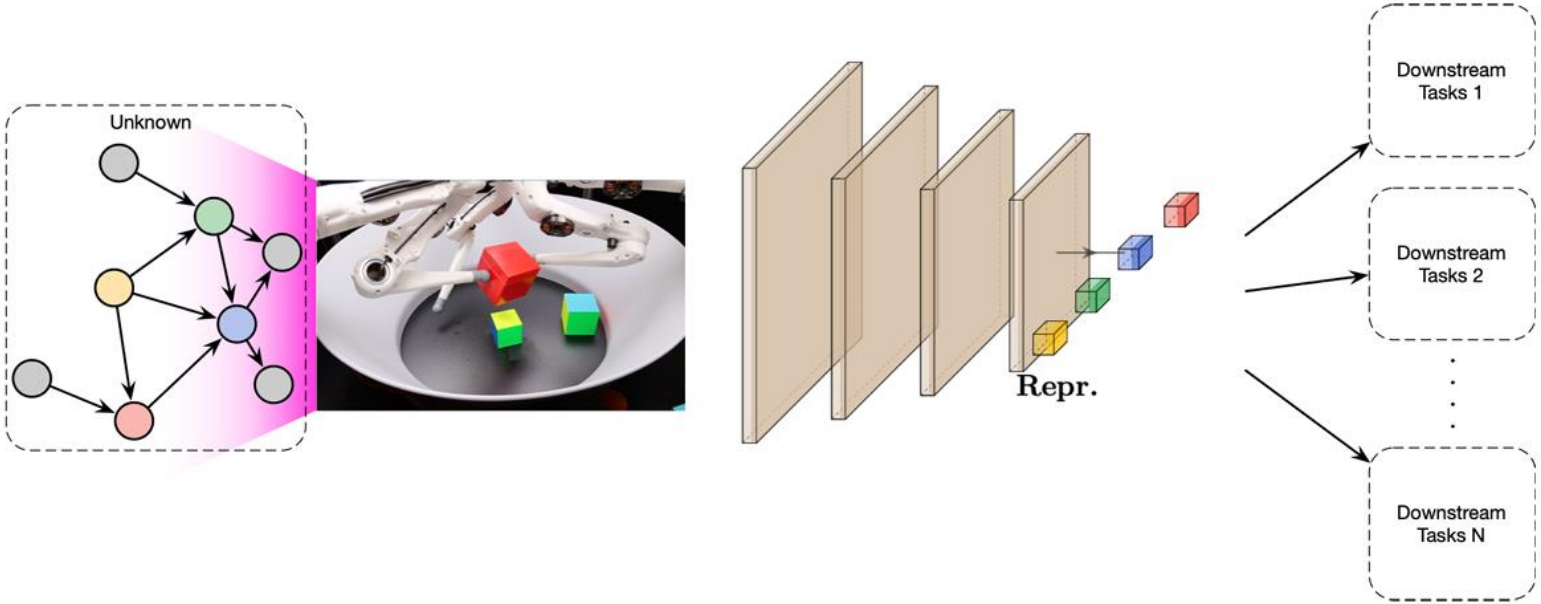
here, g is called "mixing function"

…that generates the entangled observations…

…which are also influenced by random noise

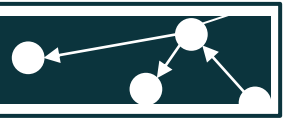
[4] Squires, Chandler. "Chapter 7 Causal Representation Learning" 6.S091-causality, 2023, https://github.com/csquires/6.S091-causality/blob/main/lecture_notes/Lecture7.pdf

CRL Basics



Schölkopf et al. "Toward causal representation learning." Proceedings of the IEEE 2021

CRL Basics



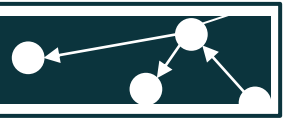
Why do we care?

- The independent mechanisms principle [5]:

> **Independent Causal Mechanisms (ICM) Principle.**
> *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

[5] Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

CRL Basics



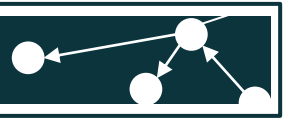
Why do we care?

- The independent mechanisms principle [5]:

> **Independent Causal Mechanisms (ICM) Principle.** *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

[5] Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

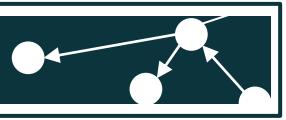
CRL Basics



Why do we care?

- The modularity implied by the independent mechanism principle allows for distribution shifts of single variables (conditional probability distributions) and interventions → going beyond the i.i.d. setting
- Requires **disentangled representation**
- Other benefits include typical advantages of causal models, such as robustness, transferability, interpretability, sample-efficiency, …

CRL Basics



Why do we care?

- The sparse mechanism shift [5]:

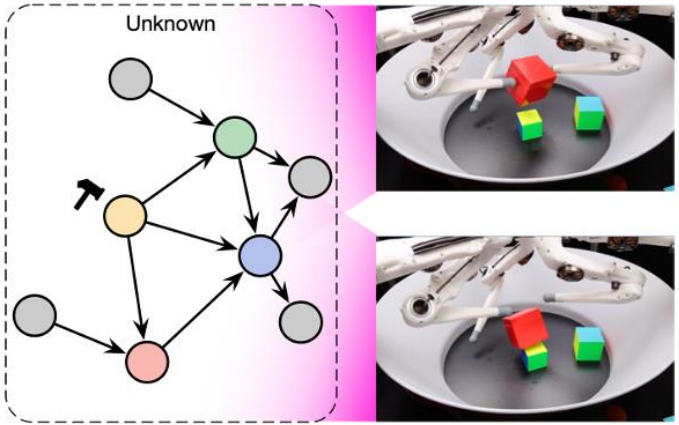
> **Sparse Mechanism Shift (SMS).** *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization (4), i.e., they should usually not affect all factors simultaneously.*

[5] Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

CRL Basics

Why do we care?

- The sparse mechanism shift [5]:



[5] cite

CRL Basics

Without any assumptions, identifiability is impossible on purely observational data. There are three basic approaches [4]

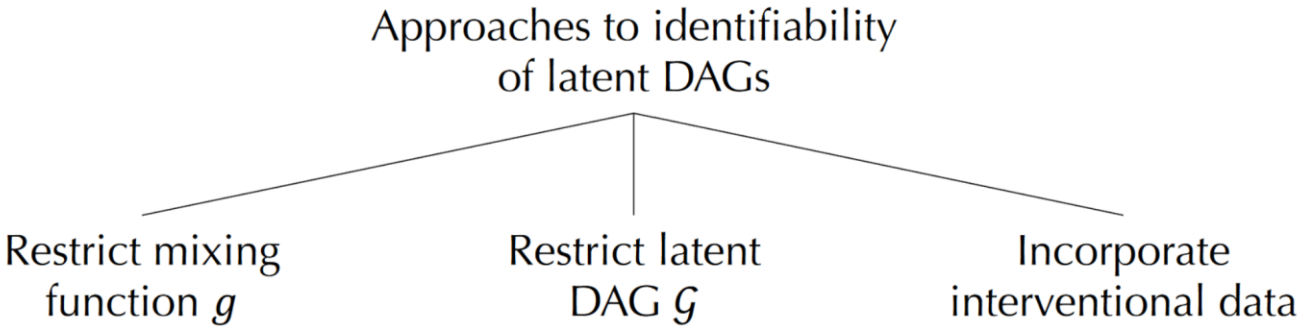


Figure 7.2: Approaches to identifiability of latent DAG models.

[4] Squires, Chandler. “Chapter 7 Causal Representation Learning” 6.S091-causality, 2023, https://github.com/csquires/6.S091-causality/blob/main/lecture_notes/Lecture7.pdf

CRL Basics

Without any assumptions, identifiability is impossible on purely observational data. There are three basic approaches

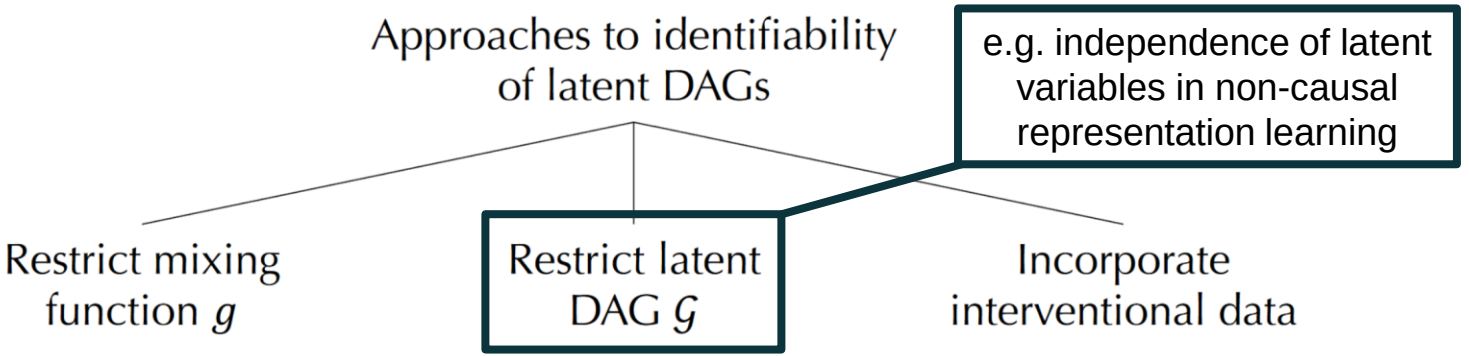
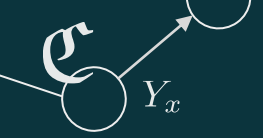
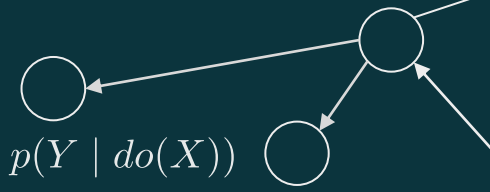


Figure 7.2: Approaches to identifiability of latent DAG models.

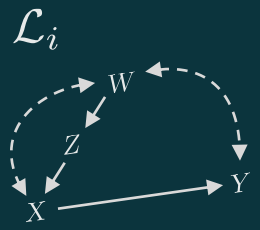
[4] Squires, Chandler. “Chapter 7 Causal Representation Learning” 6.S091-causality, 2023, https://github.com/csquires/6.S091-causality/blob/main/lecture_notes/Lecture7.pdf





Section

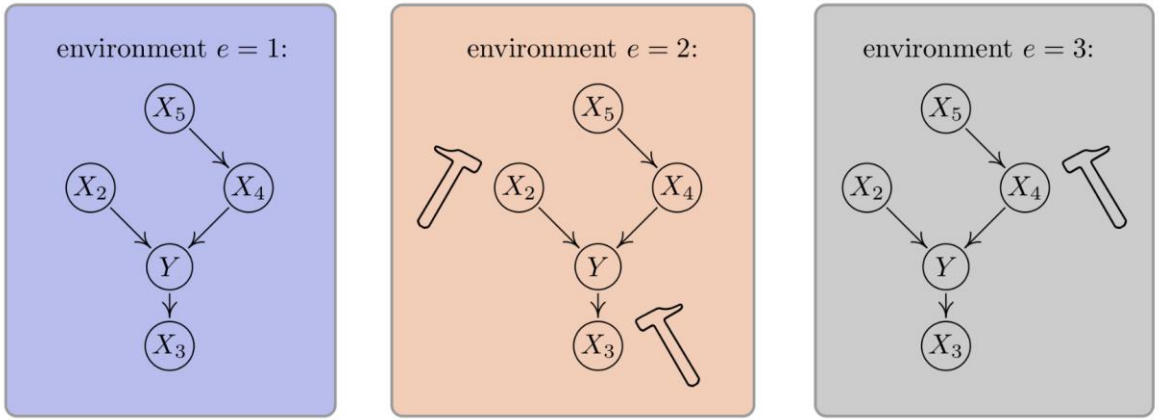
4 (Some) Causal Representation Learning Approaches



Invariant Causal Prediction

Invariant Causal Prediction (ICP)

- Given data from different environments (i.e. interventions)

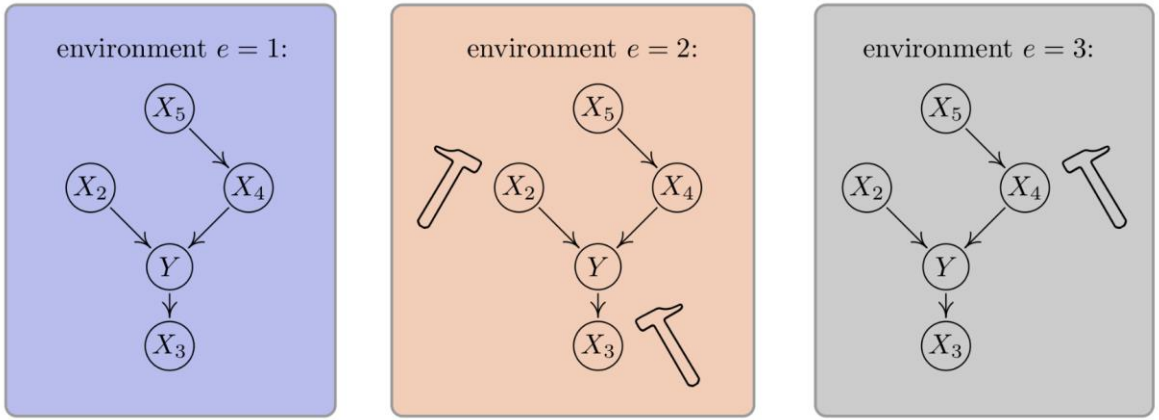


Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2016

Invariant Causal Prediction

Invariant Causal Prediction (ICP)

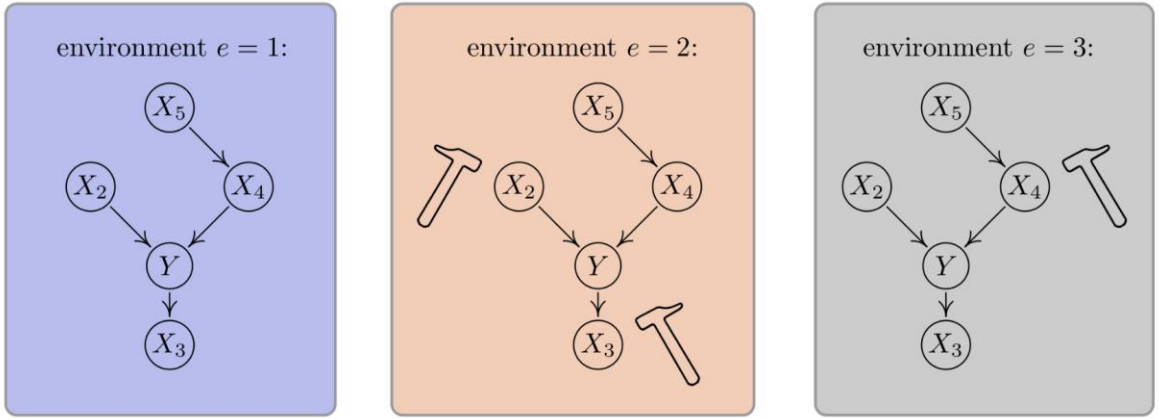
- Given data from different environments (i.e. interventions)
- Goal: Classify Y



Invariant Causal Prediction

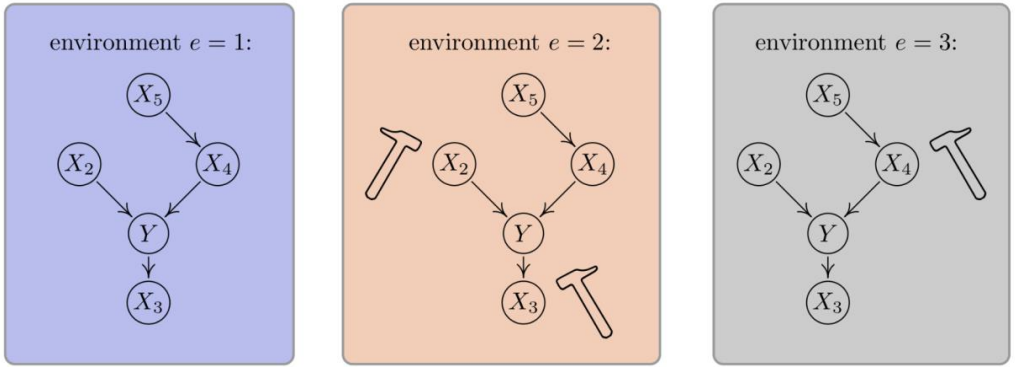
Invariant Causal Prediction (ICP)

- Given data from different environments (i.e. interventions)
- Goal: Classify Y
- Idea: Structural assignment for Y remains identical for interventions not on Y

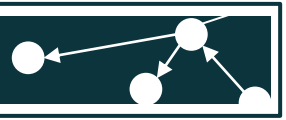


Invariant Causal Prediction

- identify causal relationships
- construct valid confidence intervals for the effects of interventions
- Even when model is mis-specified, ICP can identify a subset of causal variables by ensuring that the predictions remain invariant across envs.



Invariant Causal Prediction



Given: Datasets $D_e := (X^e, Y^e)$ from environments $e \in \mathcal{E}$

Assumption 1 (Invariant prediction) *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies*

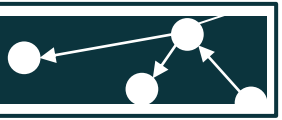
$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$
$$Y^e = \mu + X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \tag{3}$$

where $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance and the same distribution* F_ε *across all* $e \in \mathcal{E}$.

S*: set of predictors
X^e: predictor variable
Y^e: target variable

- Causal relationships remain invariant across different environments
- If a set of variables can predict the target variable in all environments, it is likely to contain only causal variables
- Interested in settings where such careful experimentation is not possible
- different distributions of X^e in the environments are generated by unknown and not precisely controlled interventions

Invariant Causal Prediction [6]



Given: Datasets $D_e := (X^e, Y^e)$ from environments $e \in \mathcal{E}$

Assumption 1 (Invariant prediction) *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies*

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$
$$Y^e = \mu + X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \tag{3}$$

where $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance and the same distribution* F_ε *across all* $e \in \mathcal{E}$.

→ Linear functions and no unobserved confounders

- only present results for the linear Gaussian models
- Assumption: the intervention does not change the conditional distribution of the target given the causal predictors

- not specific to representation learning -

On the linearity assumption in causality

Reasons and implications of the linearity assumption

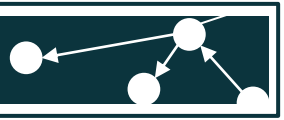
- not specific to representation learning -

On the linearity assumption in causality

Reasons and implications of the linearity assumption

- Restricting the space of functions rules out many possibilities, making **identification easier** (or even possible)
- It is one of the (if not the) most **common** function appearing in our world
- “Knowing” (by assumption) the function, it is possible to **extrapolate** outside of the training data (also true for other functions)
- Even if a more general function could be learned by a general function approximator (e.g. neural network), extrapolation might fail completely

iCaRL



invariant Causal Representation Learning (iCaRL)

Lu, Chaochao, et al. "Nonlinear invariant risk minimization: A causal approach." arXiv preprint arXiv:2102.12353 (2021).

iCaRL [7]

Not to be confused with: Incremental Classifier and Representation Learning (iCaRL)

invariant Causal Representation Learning (iCaRL)

- leverages causal inference principles to identify and utilize invariant features across environments

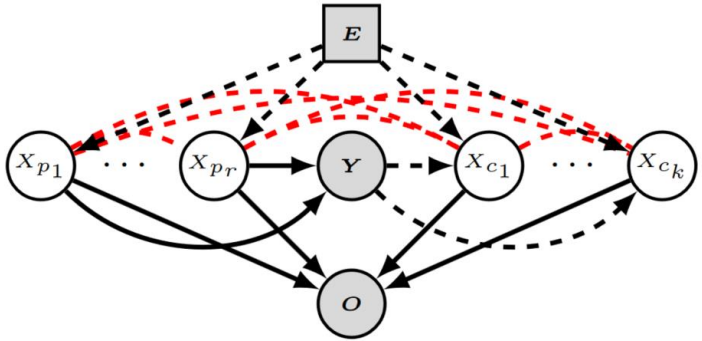
[7] Lu, Chaochao, et al. "Nonlinear invariant risk minimization: A causal approach." arXiv preprint arXiv:2102.12353 (2021).

iCaRL

Not to be confused with: Incremental Classifier and Representation Learning (iCaRL)

invariant Causal Representation Learning (iCaRL)

- Same goal as ICP but allows for non-linear functions
- Latent variables X can be connected in any way that results in a DAG



iCaRL

Not to be confused with: Incremental Classifier and Representation Learning (iCaRL)

invariant Causal Representation Learning (iCaRL)

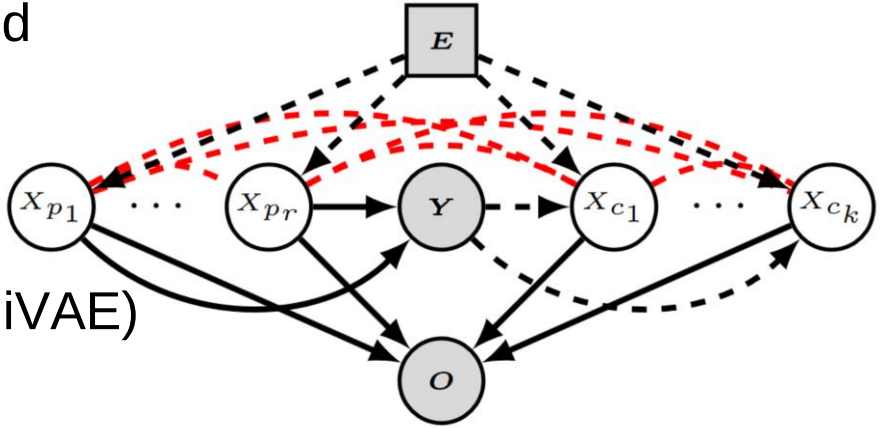
- optimization framework for training models under the Invariant Risk Minimization (IRM) approach
- define a loss function that penalizes the variability of the risk across different environments

iCaRL [7]

Not to be confused with: Incremental Classifier and Representation Learning (iCaRL)

invariant Causal Representation Learning (iCaRL)

- Same goal as ICP but allows for non-linear functions
- Latent variables X can be connected in any way that results in a DAG
- Three steps:



1. Identify latent variables (extended iVAE)
2. Determine direct causes of Y (PC)
3. Learn invariant predictor based on direct causes only

iCaRL

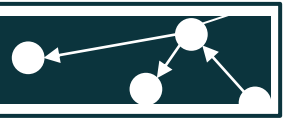
Experiment on colored MNST

- Color is spuriously correlated with label (digit)

Top row: Intervention on cause variable changes shape but not color

Bottom row: Intervention on effect variable changes color but not shape

CausalVAE



CausalVAE

- Variational autoencoder that encodes an SCM in the latent representation

Yang et al. "Causalvae: Disentangled representation learning via neural structural causal models." CVPR 2021.

CausalVAE

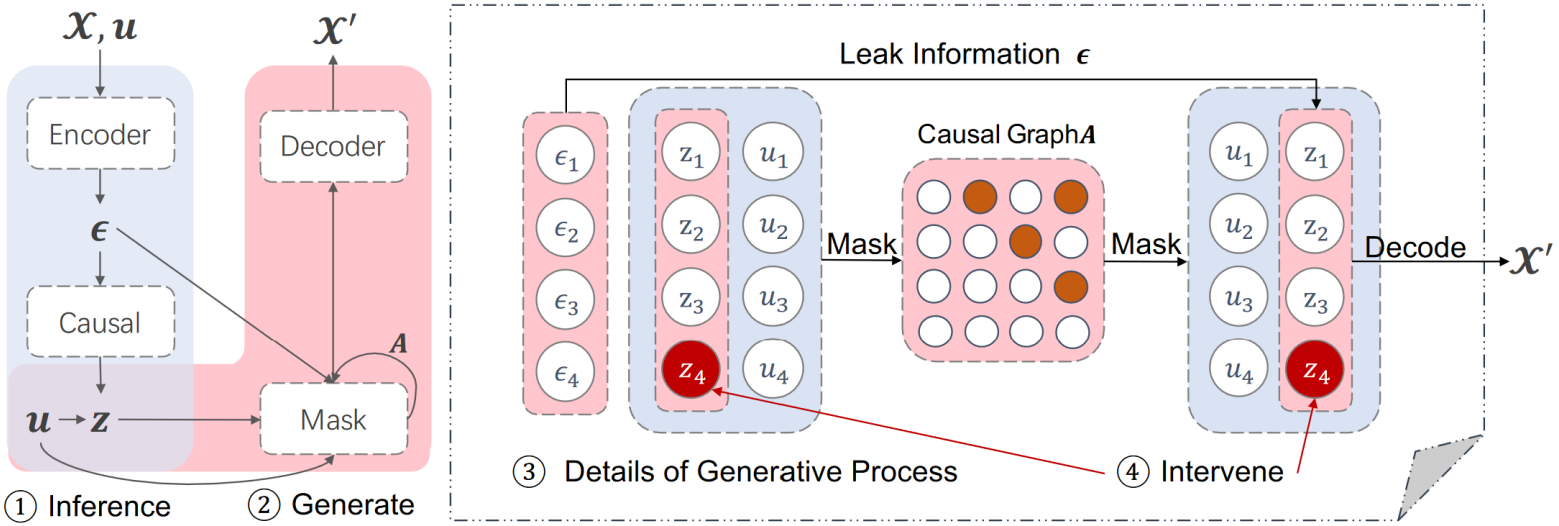
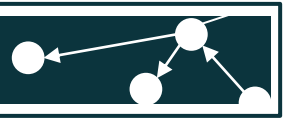


Figure 2. Model structure of CausalVAE. The encoder takes observation $\mathbf{x}$ as inputs to generate independent exogenous variable $\boldsymbol{\epsilon}$, whose prior distribution is assumed to be standard Multivariate Gaussian. Then it is transformed by the Causal Layer into causal representations $\mathbf{z}$ (Eq. 1) with a conditional prior distribution $p(\mathbf{z}|\mathbf{u})$. A Mask Layer is then applied to $\mathbf{z}$ to resemble the SCM in Eq. 2. After that, $\mathbf{z}$ is taken as the input of the decoder to reconstruct the observation $\mathbf{x}$.

Eq. 1: $z = A^T z + \epsilon = (I - A^T)^{-1}\epsilon, \epsilon \sim \mathcal{N}(0, I)$

Eq. 2: $\mathrm{z_i} = g_i(A_i \circ z; \eta_i) + \epsilon_i$

CausalVAE



CausalVAE

- Variational autoencoder that encodes an SCM in the latent representation
- True causal concepts u given during training
- Masking layer with adjacency matrix A allows for interventions
- Loss includes acyclicity constraint on A

$$\mathcal{L} = -\text{ELBO} + \alpha H(\mathbf{A}) + \beta l_u + \gamma l_m$$

$$H(\mathbf{A}) \equiv tr((\mathbf{I} + \frac{c}{m}\mathbf{A} \circ \mathbf{A})^n) - n = 0$$

$$l_u = \mathbb{E}_{q_\mathcal{X}} \|\mathbf{u} - \sigma(\mathbf{A}^T \mathbf{u})\|_2^2 \leq \kappa_1$$

$$l_m = \mathbb{E}_{\mathbf{z} \sim q_\Phi} \sum_{i=1}^{n} \|z_i - g_i(\mathbf{A}_i \circ \mathbf{z}; \boldsymbol{\eta}_i)\|^2 \leq \kappa_2$$

CausalVAE [8]

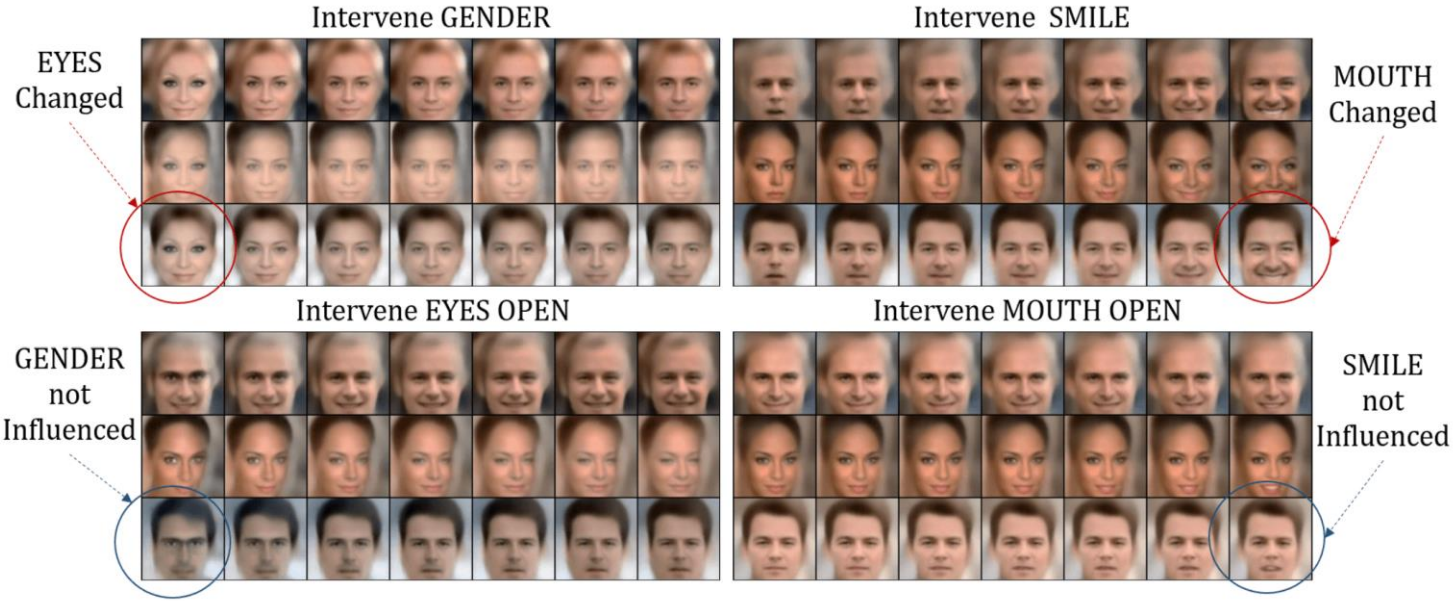
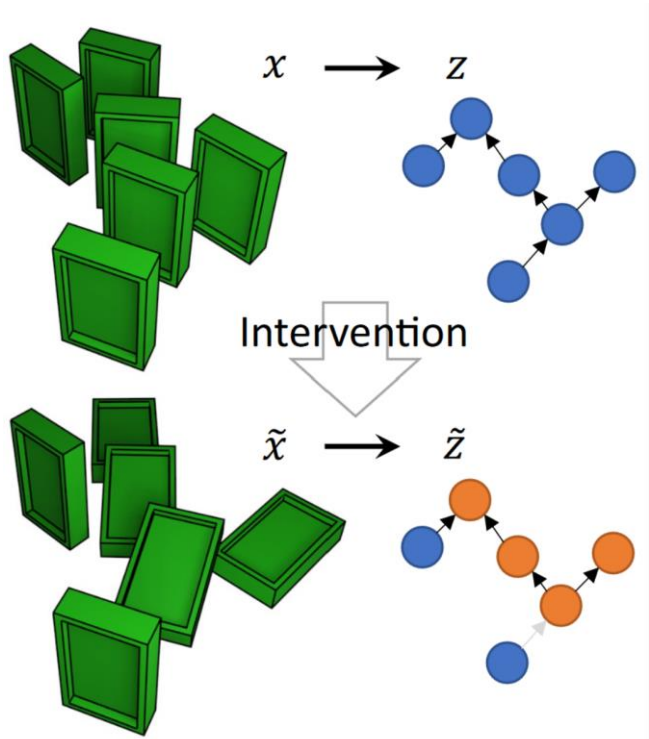


Figure 4. Results of CausalVAE model on CelebA(SMILE). The controlled factors are GENDER, SMILE, EYES OPEN and MOUTH OPEN respectively. More intervention results are shown in Appendix D.3.

ILCM

Implicit latent causal models (ILCMs)

- Learn causal representations from pixels by using pairs of (counterfactual) samples x and $\tilde{x}$; no other labels
- Need to observe an intervention on any variable that should be identified



Brehmer, Johann, et al. "Weakly supervised causal representation learning." NeurIPS 2022

ILCM [9]

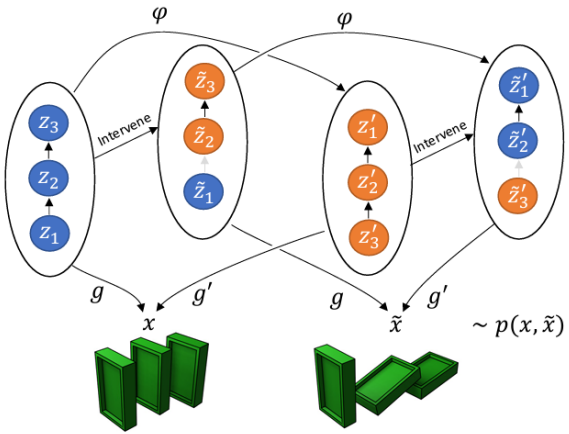


Figure 2: In LCM $\mathcal{M}$, z_i denotes whether the i-th stone from the front is standing. Intervening on the second variable, z_2, leads to $\tilde{z}$. The decoder g renders $z, \tilde{z}$ as images $x, \tilde{x}$. LCM $\mathcal{M}'$ has an equivalent representation in which z'_i denotes whether the i-th stone from the *back* has *fallen*. In Thm. 1, we prove that if and only if two causal models have the same pixel distribution $p(x, \tilde{x})$, there exists an LCM isomorphism φ: an element-wise reparameterization of the causal variables plus a permutation of the ordering that commutes with interventions and causal mechanisms.

Definition 1 (Latent causal model (LCM)). *A latent causal model $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ consists of*

- *an acyclic SCM $\mathcal{C}$, which is faithful (all independencies are encoded in its graph [24]),*
- *an observation space $\mathcal{X}$,*
- *a decoder $g : \mathcal{Z} \to \mathcal{X}$ that is diffeomorphic onto its image,*
- *a set $\mathcal{I}$ of interventions on $\mathcal{C}$, and*
- *a probability measure $p_{\mathcal{I}}$ over $\mathcal{I}$.*

Theorem 1 (Identifiability of $\mathbb{R}$-valued LCMs from weak supervision). *Let $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{M}' = \langle \mathcal{C}', \mathcal{X}, g', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be LCMs with the following properties:*

- *The LCMs have an identical observation space $\mathcal{X}$.*
- *The SCMs $\mathcal{C}$ and $\mathcal{C}'$ both consist of n real-valued endogeneous causal variables and corresponding exogenous noise variables, i. e. $\mathcal{E}_i = \mathcal{Z}_i = \mathcal{Z}'_i = \mathcal{E}'_i = \mathbb{R}$.*
- *The intervention sets $\mathcal{I}$ and $\mathcal{I}'$ consist of all atomic, perfect interventions, $\mathcal{I} = \{\emptyset, \{z_0\}, \ldots, \{z_n\}\}$ and similar for $\mathcal{I}'$.*
- *The intervention distribution $p_{\mathcal{I}}$ and $p'_{\mathcal{I}'}$ have full support.*

Then the following two statements are equivalent:

1. *The LCMs entail equal weakly supervised distributions, $p^{\mathcal{X}}_{\mathcal{M}}(x, \tilde{x}) = p^{\mathcal{X}}_{\mathcal{M}'}(x, \tilde{x})$.*
2. *The LCMs are equivalent in the sense of Def. 2, $\mathcal{M} \sim \mathcal{M}'$.*

ILCM [9]

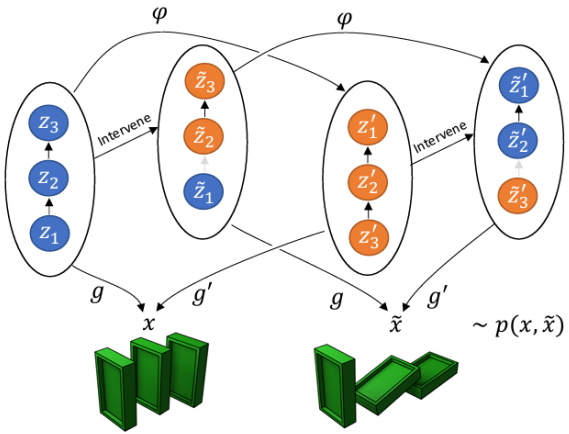


Figure 2: In LCM $\mathcal{M}$, z_i denotes whether the i-th stone from the front is standing. Intervening on the second variable, z_2, leads to $\tilde{z}$. The decoder g renders $z, \tilde{z}$ as images $x, \tilde{x}$. LCM $\mathcal{M}'$ has an equivalent representation in which z'_i denotes whether the i-th stone from the *back* has *fallen*. In Thm. 1, we prove that if and only if two causal models have the same pixel distribution $p(x, \tilde{x})$, there exists an LCM isomorphism φ: an element-wise reparameterization of the causal variables plus a permutation of the ordering that commutes with interventions and causal mechanisms.

Definition 1 (Latent causal model (LCM)). *A latent causal model $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ consists of*

- *an acyclic SCM $\mathcal{C}$, which is faithful (all independencies are encoded in its graph [24]),*
- *an observation space $\mathcal{X}$,*
- *a decoder $g : \mathcal{Z} \to \mathcal{X}$ that is diffeomorphic onto its image,*
- *a set $\mathcal{I}$ of interventions on $\mathcal{C}$, and*
- *a probability measure $p_{\mathcal{I}}$ over $\mathcal{I}$.*

Theorem 1 (Identifiability of $\mathbb{R}$-valued LCMs from weak supervision). *Let $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{M}' = \langle \mathcal{C}', \mathcal{X}, g', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be LCMs with the following properties:*

- *The LCMs have an identical observation space $\mathcal{X}$.*
- *The SCMs $\mathcal{C}$ and $\mathcal{C}'$ both consist of n real-valued endogeneous causal variables and corresponding exogenous noise variables, i. e. $\mathcal{E}_i = \mathcal{Z}_i = \mathcal{Z}'_i = \mathcal{E}'_i = \mathbb{R}$.*
- *The intervention sets $\mathcal{I}$ and $\mathcal{I}'$ consist of all atomic, perfect interventions, $\mathcal{I} = \{\emptyset, \{z_0\}, \ldots, \{z_n\}\}$ and similar for $\mathcal{I}'$.*
- *The intervention distribution $p_{\mathcal{I}}$ and $p'_{\mathcal{I}'}$ have full support.*

Then the following two statements are equivalent:

1. *The LCMs entail equal weakly supervised distributions, $p^{\mathcal{X}}_{\mathcal{M}}(x, \tilde{x}) = p^{\mathcal{X}}_{\mathcal{M}'}(x, \tilde{x})$.*
2. *The LCMs are equivalent in the sense of Def. 2, $\mathcal{M} \sim \mathcal{M}'$.*

Main result: an LCM M can be identified from p(x, ˜ x) up to a relabeling and elementwise transformations of the causal variables

ILCM

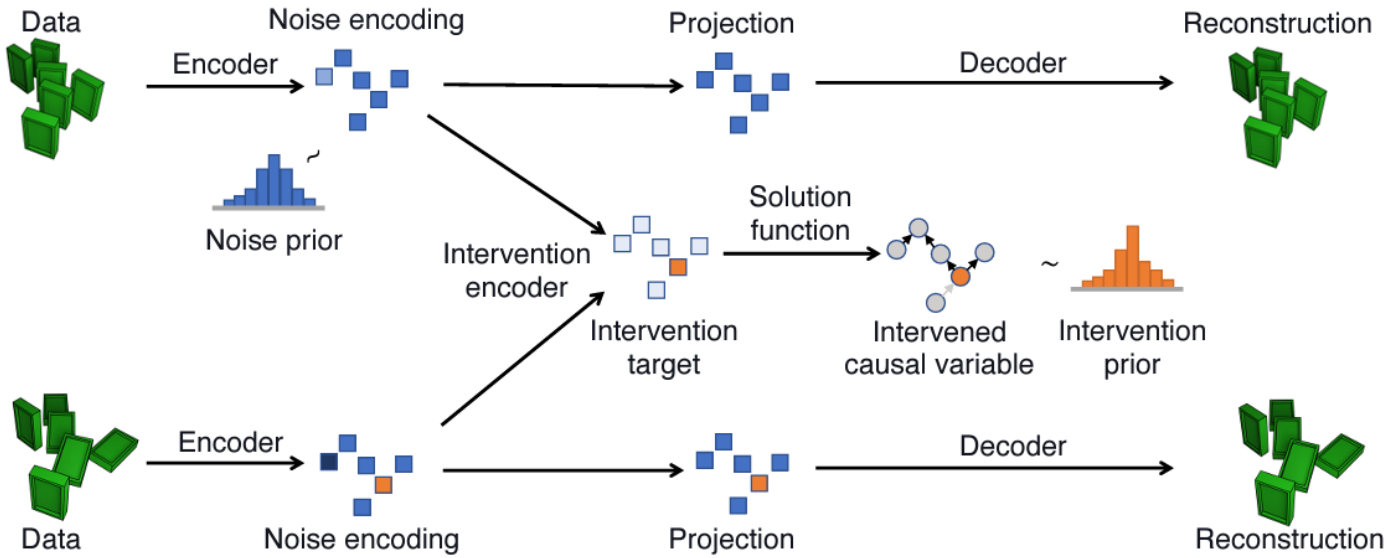


Figure 3: ILCM architecture. Pre- and post-intervention data (left) are encoded to noise encodings and intervention targets, which are then decoded back to the data space. To compute the prior probability density, the noise encodings are transformed into causal variables with the neural solution function.

ILCM

Nice results on synthetic datasets

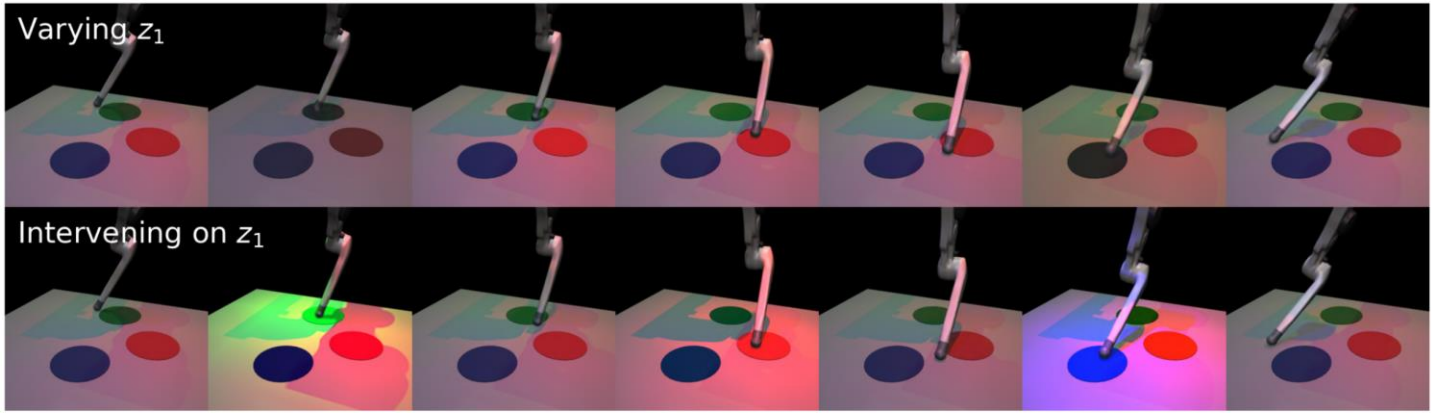


Figure 6: Varying learned causal factors vs. intervening on them. With a trained ILCM, we encode a single test image (left column). In the top row, we then vary the latent z_1 independently, without computing causal effects, and show the corresponding reconstructed images. Only the robot arm position changes, highlighting that we learned a disentangled representation. In the bottom row we instead *intervene* on z_1 and observe the causal effects: the robot arm may activate lights, which in turn can affect other lights in the circuit.

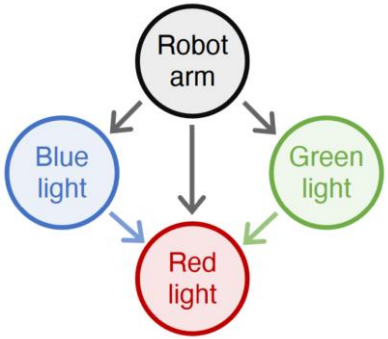
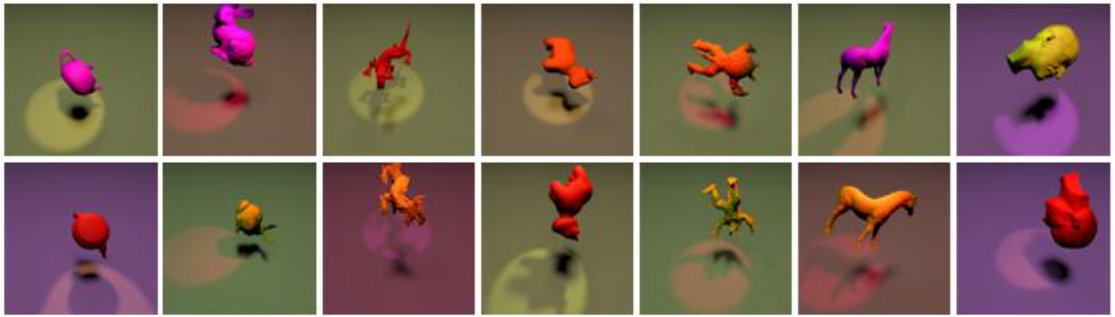


Figure 7: Causal graph of the CausalCircuit dataset.

Content-Style Separation

A causal view on data augmentation

- Goal: learn separate representations for content and style of images



Von Kügelgen, Julius, et al. "Self-supervised learning with data augmentations provably isolates content from style." NeurIPS 2021

Content-Style Separation

A causal view on data augmentation

- Goal: learn separate representations for content and style of images
- Observation x, content vars c, style vars s, and augmentations $\tilde{s}$ and $\tilde{x}$

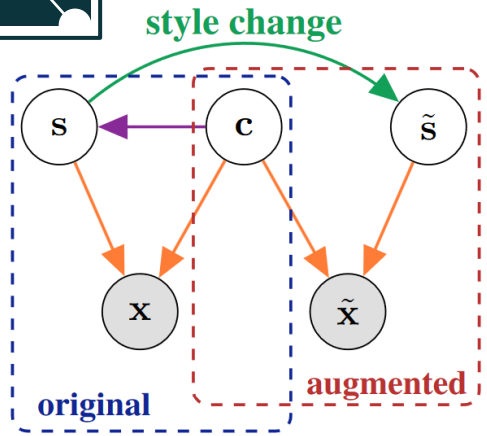


Figure 1: **Overview of our problem formulation.** We partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, and allow for statistical and causal dependence of style on content. We assume that only style changes between the original view $\mathbf{x}$ and the augmented view $\tilde{\mathbf{x}}$, i.e., they are obtained by applying the same deterministic function $\mathbf{f}$ to $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ and $\tilde{\mathbf{z}} = (\mathbf{c}, \tilde{\mathbf{s}})$.

Content-Style Separation

A causal view on data augmentation

- Goal: learn separate representations for content and style of images
- Observation x, content vars c, style vars s, and augmentations $\tilde{s}$ and $\tilde{x}$
- **Why $c \rightarrow s$ and not $s \rightarrow c$?**

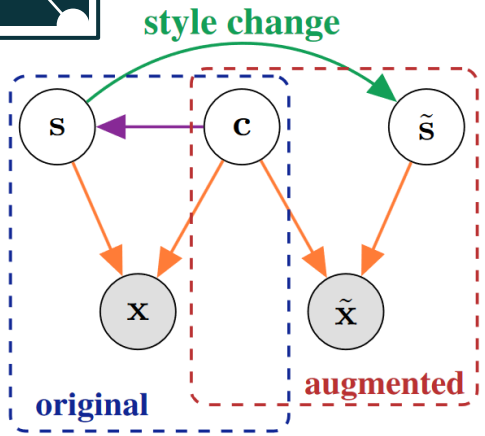


Figure 1: **Overview of our problem formulation.** We partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, and allow for statistical and causal dependence of style on content. We assume that only style changes between the original view $\mathbf{x}$ and the augmented view $\tilde{\mathbf{x}}$, i.e., they are obtained by applying the same deterministic function $\mathbf{f}$ to $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ and $\tilde{\mathbf{z}} = (\mathbf{c}, \tilde{\mathbf{s}})$.

Content-Style Separation [10]

A causal view on data augmentation

- Goal: learn separate representations for content and style of images
- Observation x, content vars c, style vars s, and augmentations $\tilde{s}$ and $\tilde{x}$
- Why $c \rightarrow s$ and not $s \rightarrow c$? Because the class prediction should be invariant to the style, it must not depend on it

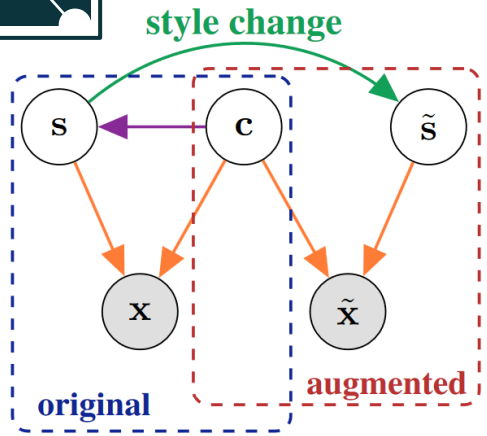
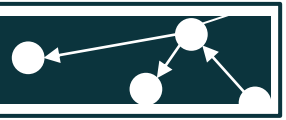


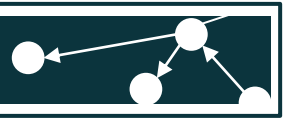
Figure 1: **Overview of our problem formulation.** We partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, and allow for statistical and causal dependence of style on content. We assume that only style changes between the original view $\mathbf{x}$ and the augmented view $\tilde{\mathbf{x}}$, i.e., they are obtained by applying the same deterministic function $\mathbf{f}$ to $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ and $\tilde{\mathbf{z}} = (\mathbf{c}, \tilde{\mathbf{s}})$.

Content-Style Separation



- Counterfactual question: *“what would have happened if the style variables had been (randomly) perturbed, all else being equal?”*
- Given a fixed size of the content representation, identification can be achieved by finding a function with the same (or very similar) outputs on the original and augmented images x and $\tilde{x}$ while avoiding that a collapsed representation is learned

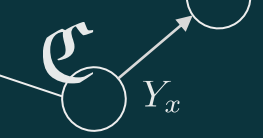
Content-Style Separation [10]

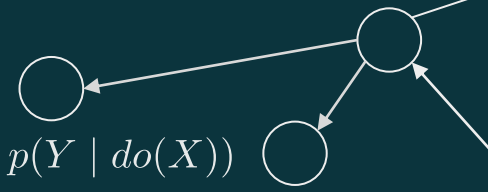


Possible loss function:

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left\| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right\|_2^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right)$$

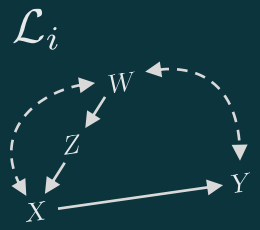
Here, the differential entropy H avoids a collapsed representation



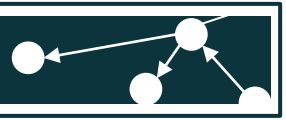


Section

5 Summary & Conclusion

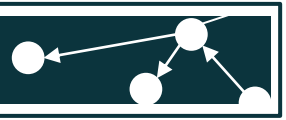


Summarizing Thoughts



What works well? What are problems and challenges?

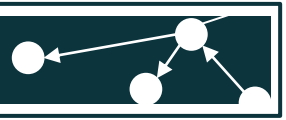
Summarizing Thoughts



What works well? What are problems and challenges?

- Given a specific observation, there is not just one single “true” representation
- Current research is often tested on synthetic problems and relies on specific assumptions

Summarizing Thoughts

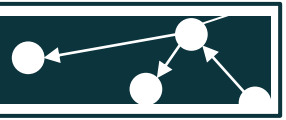


What works well? What are problems and challenges?

- Given a specific observation, there is not just one single “true” representation
- Current research is often tested on synthetic problems and relies on specific assumptions

Lack of benchmarks!

Summarizing Thoughts

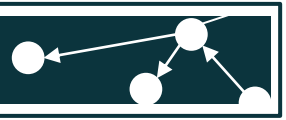


What works well? What are problems and challenges?

- Given a specific observation, there is not just one single "true" representation
- Current research is often tested on synthetic problems and relies on specific assumptions
- If these assumptions are satisfied, representation learning and manipulation of the latent space works well
- Still an open research topic

Lack of benchmarks!

Conclusion



- Causal representation learning: learning low-dimensional representations of higher-dimensional data where the latent representation is an SCM
- Causal representations have several advantages due to modularity
- Identification is impossible without assumptions or further information (i.e. interventional or counterfactual data)
- Current research mostly focuses specific problems