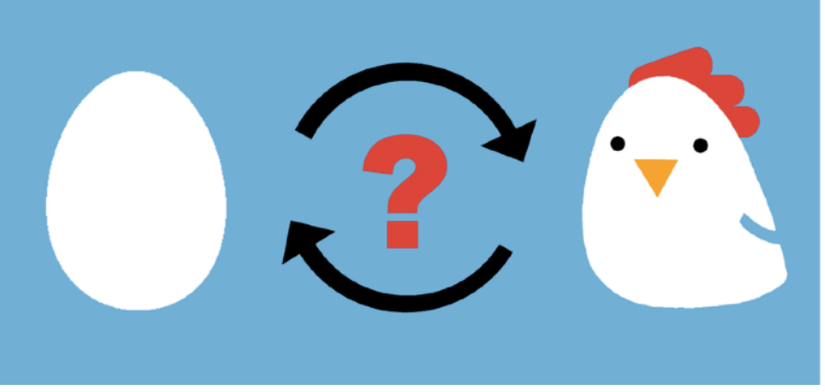


Causality and Machine Learning
(80-816/516)

Classes 19 (March 25, 2025)

Causal Representation Learning 2: Benefits from multiple distributions (from changes)

Instructor:

Kun Zhang (kunz1@cmu.edu)

Zoom link: https://cmu.zoom.us/j/8214572323)

Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

We Mainly Focused on the IID Case: Recent Advances in Causal Representation Learning

i.i.d. data?	Parametric constraints?	Latent confounders?	What can we get?	
Yes	*No*	*No*	*(Different types of) equivalence class*	- PC, FCI, etc.
		Yes		
	Yes	**No**	**Unique identifiability (under structural conditions)**	- LiNGAM - Rank-based, GIN…
		Yes		
Non-I, but I.D.	No/Yes	No	(Extended) regression	
		Yes	Latent temporal causal processes identifiable!	
I., but non-I.D.	No	No	More informative than MEC (CD-NOD)	- CD-NOD
	Yes		May have unique identifiability	
	No	Yes	Changing subspace identifiable	- CRL from multiple distributions
	Yes		Variables in changing relations identifiable	- Causal GenAI

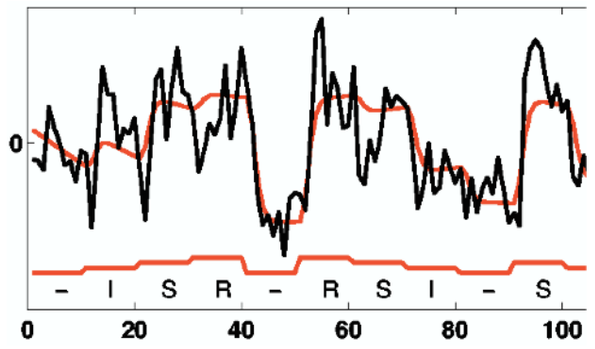
CRL from Changes: Outline

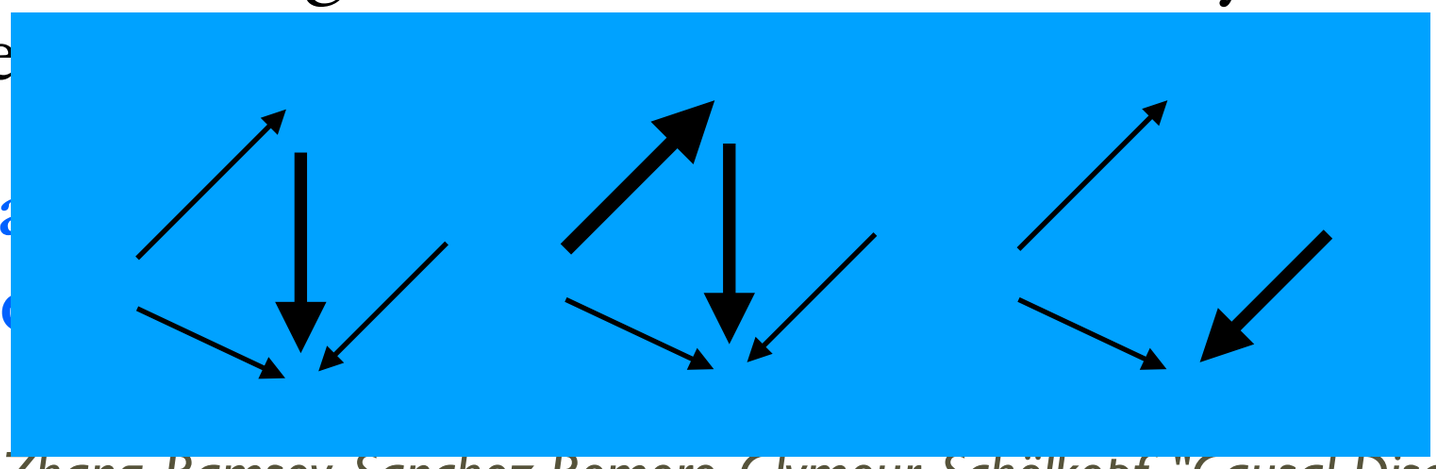
I., but non-I.D.	No	No	More informative than MEC (CD-NOD)
	Yes		May have unique identifiability
	No	Yes	Changing subspace identifiable
	Yes		Variables in changing relations identifiable

- CD-NOD (Causal Discovery from Nonstationary/Heterogenuous data)
- Nonlinear ICA with partial changes across domains
  - Partial disentanglement
  - Domain adaptation, image translation, and multi-domain data generation
  - Learning from text-image pairs
- A general setting
- Connection to the IID case: Synergy between minimal changes and sparsity

Nonstationary/Heterogeneous Data and Causal Modeling

- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data...)
  - Multiple data sets under different observational or experimental conditions
- Causal modeling & distribution shift heavily couple
  - $P(c$ inde





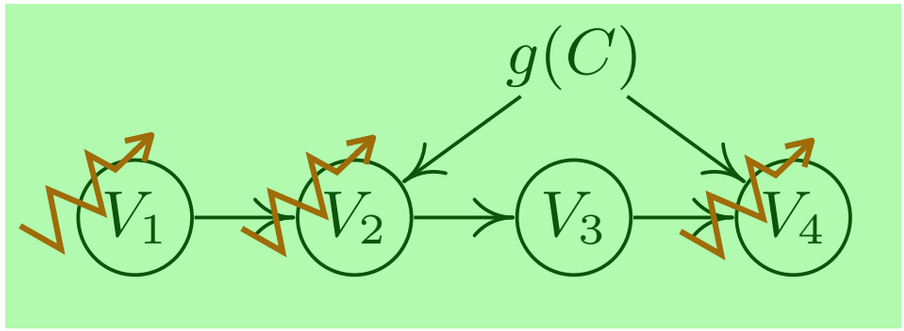
Huang, Zhang, Zhang, Ramsey, Sanchez-Romero, Glymour, Schölkopf, "Causal Discovery from Heterogeneous/ Nonstationary Data," JMLR, 2020

Zhang, Huang, et al., Discovery and visualization of nonstationary causal models, arxiv 2015

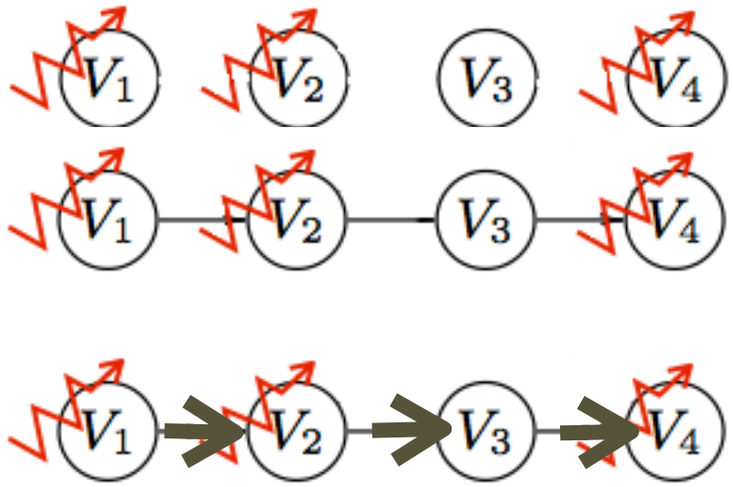
Ghassami, et al., Multi-Domain Causal Structure Learning in Linear Systems, NIPS 2018

Causal Discovery from Nonstationary/ Heterogeneous Data

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	*Yes*



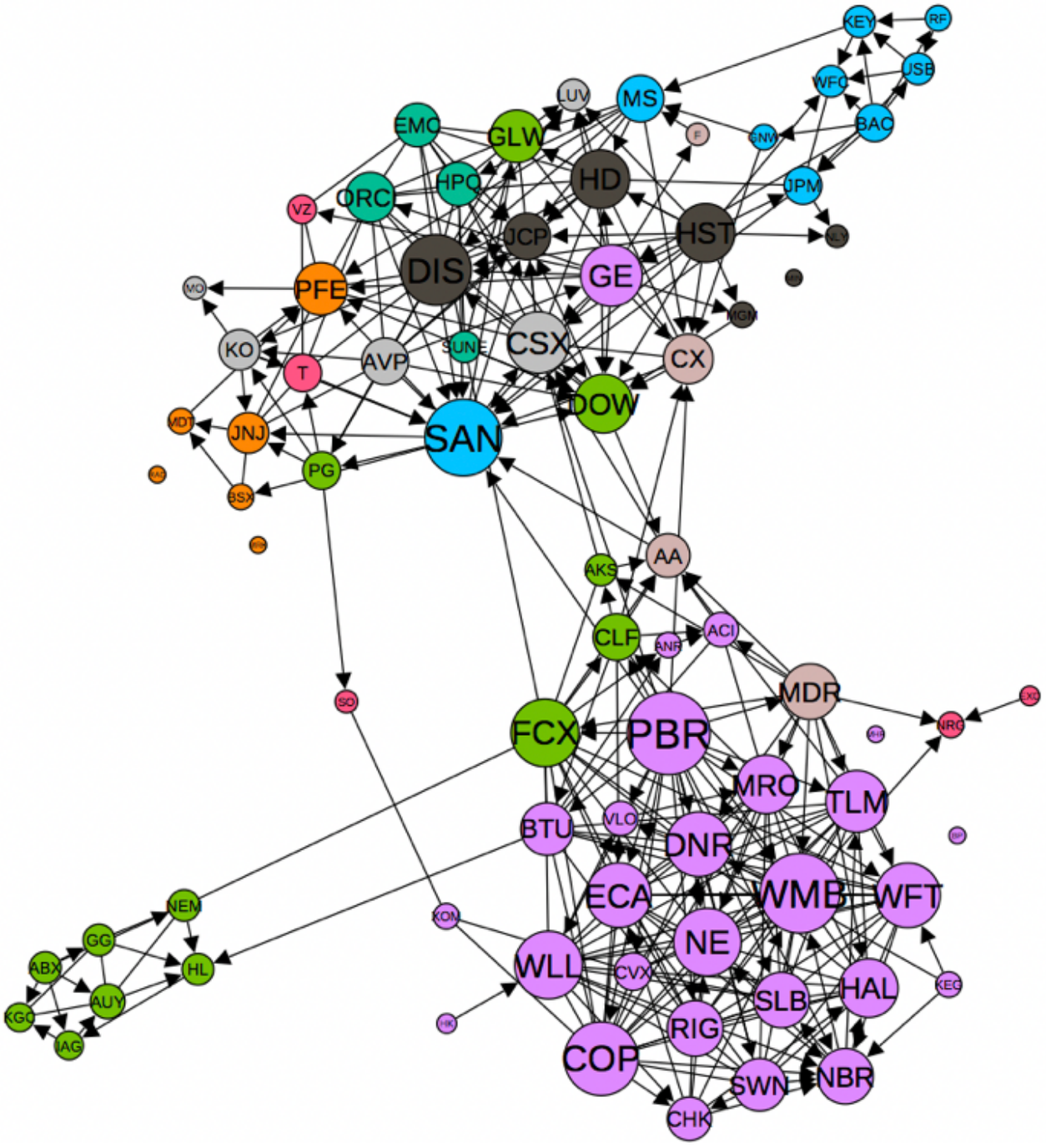
- Task:
  - Determine changing causal modules & estimate skeleton
- Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause), including invariant mechanism/ cause as special cases
- Visualization of changing modules over time/ across data sets?

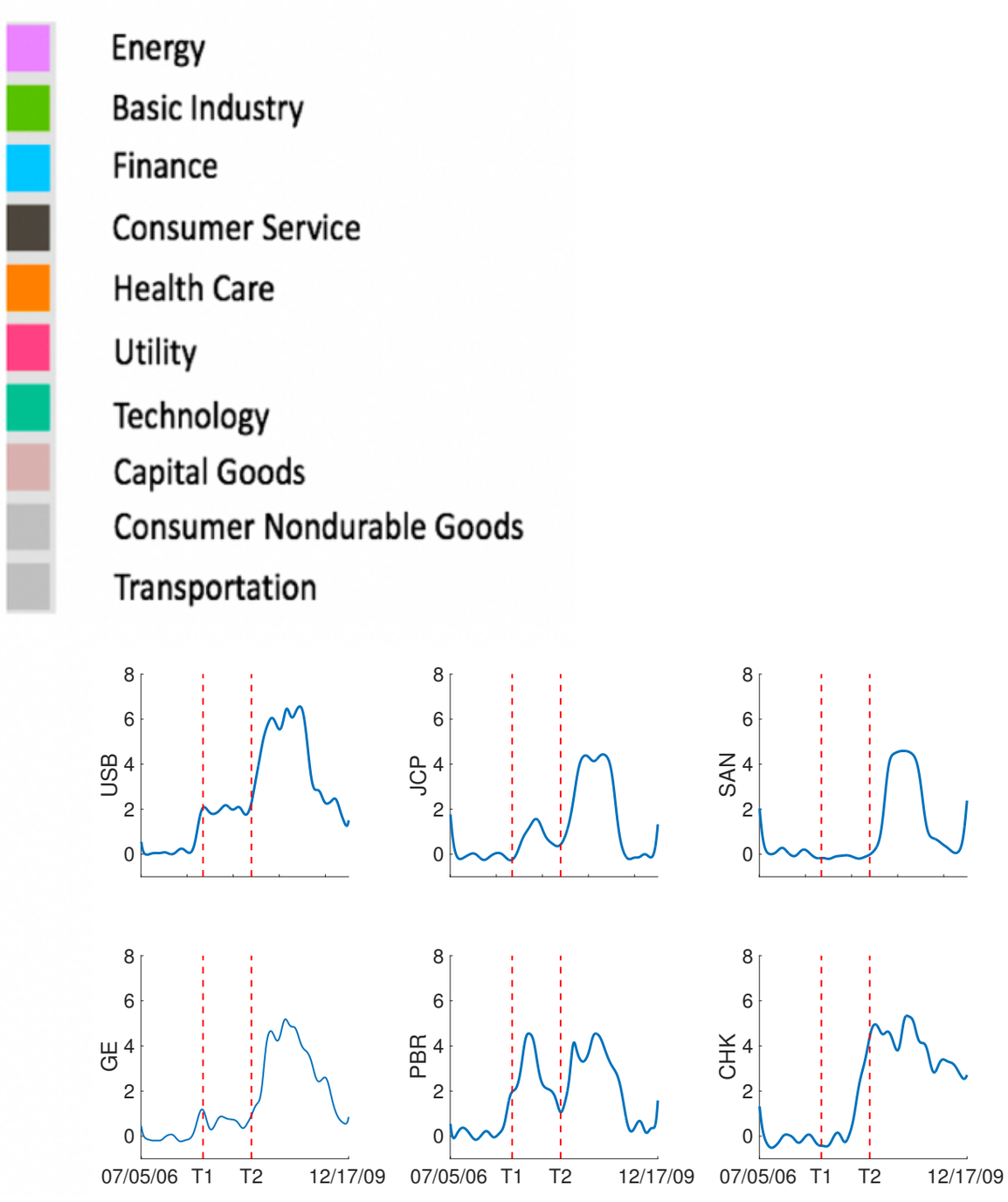


Kernel nonstationary driving force estimation

- *Huang et al., "Causal Discovery from Heterogeneous/Nonstationary Data," JMLR, 2020*
- *Tian, Pearl, "Causal discovery from changes," UAI 2001*
- *Hoover, "The logic of causal inference" Economics and Philosophy, 6:207–234, 1990.*

Causal Analysis of Major Stocks in NYSE (07/05/2006 - 12/16/2009)





- Huang, Zhang, Zhang, Romero, Glymour, Schölkopf, Behind Distribution Shift: Mining Driving Forces of Changes and Causal Arrows," ICDM 2017

CRL from Changes: Outline

I., but non-I.D.	No	No	More informative than MEC (CD-NOD)
	Yes		May have unique identifiability
	No	Yes	Changing subspace identifiable
	Yes		Variables in changing relations identifiable

- CD-NOD (Causal Discovery from Nonstationary/Heterogenuous data)
- Nonlinear ICA with partial changes across domains
  - Partial disentanglement
  - Domain adaptation, image translation, and multi-domain data generation
  - Learning from text-image pairs
- A general setting
- Connection to the IID case: Synergy between minimal changes and sparsity

Nonlinear Cases Generally Non-Identifiable

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

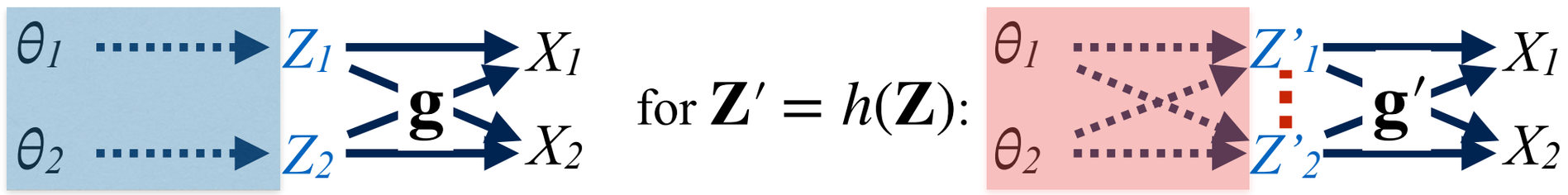
- Nonlinear ICA:
  - Generative model: $\mathbf{X} = f(\mathbf{S})$, where $\mathbf{S}$ has independent components
  - De-mixing procedure: $\mathbf{Y} = g(\mathbf{X})$, where $\mathbf{Y}$ components are as independent as possible
  - Solutions always exist and are highly non-unique: Why?

- *Aapo Hyvärinen, Petteri Pajunen, Nonlinear independent component analysis: Existence and uniqueness results, Neural Networks, 1999*

Nonlinear ICA with Multiple Domains

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

- Nonlinear ICA: observed variables follow $\mathbf{X} = \mathbf{g}(\mathbf{Z})$, in which Z_i are mutually independent
- Solutions to nonlinear ICA high non-unique
- If $p(Z_i)$ change across multiple domains, generally their are identifiable (up to component-wise transformations)
- Why?



- *Hyvärinen, Pajunen, Nonlinear independent component analysis: Existence and uniqueness results. Neural networks, 1999.*
- *Hyvarinen, Sasaki, Turner, "Nonlinear ICA using auxiliary variables and generalized contrastive learning," In The 22nd International Conference on Artificial Intelligence and Statistics, 2019.*

Nonlinear ICA with Multiple Domains: Intuition

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

- General principle: Each new domain brings more constraints than additional degrees of freedom
- Shared by many problems, such as multi-domain linear Gaussian source separation ($\mathbf{X} = \mathbf{A} \cdot \mathbf{S}$, but S_i are Gaussian with changing variances)
- Let's see why…

Remember this Story?

- Consider puerperal fever in the mid-19th century
- Two clinics used almost the same techniques but had very different mortality rates
- Semmelweis: Why?
  - *Hypothesis*: Unknown "cadaverous material" caused puerperal fever
  - Proposed *intervention*: washing hands
  - *Conflicted* with the established scientific and medical opinions of the time
  - Rejected by the medical community until years after his death, when Louis Pasteur *confirmed* the germ theory

https://amol-kulkarni.com/project/semmelweis/

Semmelweis, aged 42 in 1860, photograph by Borsos and Doctor

Partial Identifiability for Domain Adaptation

Lingjing Kong [1] **Shaoan Xie** [1] **Weiran Yao** [1] **Yujia Zheng** [1] **Guangyi Chen** [2 1] **Petar Stojanov** [3]
Victor Akinwande [1] **Kun Zhang** [2 1]

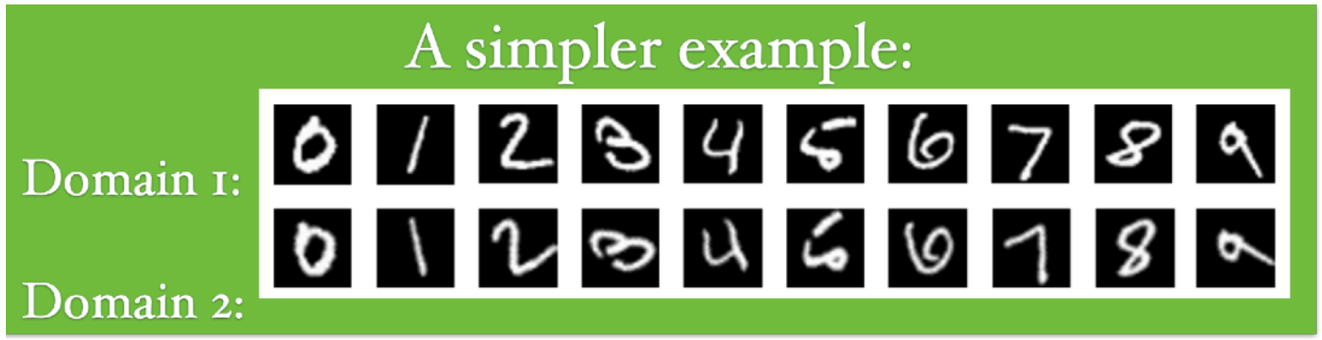
Abstract

Unsupervised domain adaptation is critical to many real-world applications where label information is unavailable in the target domain. In general, without further assumptions, the joint distribution of the features and the label is not identifiable in the target domain. To address this issue, we rely on a property of minimal changes of causal mechanisms across domains to minimize unnecessary influences of domain shift. To encode this property, we first formulate the data generating process using a latent variable model with two partitioned latent subspaces: invariant components whose distributions stay the same across domains, and sparse changing components that vary across domains. We further constrain the domain shift to have a restrictive influence on the changing components. Under mild conditions, we show that the latent variables are partially identifiable, from

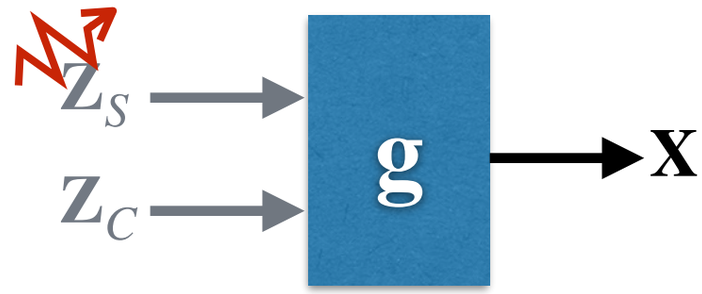
domain indices $\mathbf{u}$, the training (source domain) data follows multiple joint distributions $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_1}$, $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_2}$, ..., $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_M}$,[1] and the test (target domain) data follows the joint distribution $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$, where $p_{\mathbf{x},\mathbf{y}|\mathbf{u}}$ may vary across $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$. During training, for each i-th source domain, we are given labeled observations $(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)})_{k=1}^{m_i}$ from $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_i}$, and target domain unlabeled instances $(\mathbf{x}_k^\mathcal{T})_{k=1}^{m_\mathcal{T}}$ from $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$. The main goal of domain adaptation is to make use of the available observed information, to construct a predictor that will have optimal performance in the target domain.

It is apparent that without further assumptions, this objective is ill-posed. Namely, since the only available observations in the target domain are from the marginal distribution $p_{\mathbf{x}|\mathbf{u}^\mathcal{T}}$, the data may correspond to infinitely many joint distributions $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$. This mandates making additional assumptions on the relationship between the source and the target domain distributions, with the hope to be able to reconstruct (identify) the joint distribution in the target domain $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$. Typically, these assumptions entail some measure of sim-

Finding Changing Hidden Variables for Transfer Learning



i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



- Underlying components $\mathbf{Z}_S$ may change across domains
- Changing components $\mathbf{Z}_S$ are identifiable; invariant part $\mathbf{Z}_C$ is identifiable up to its subspace
- Using $\mathbf{Z}_C$ and transformed changing part $\tilde{\mathbf{Z}}_S$ for transfer learning

- *Kong, Xie, Yao, Zheng, Chen, Stojanov, Akinwande, Zhang, Partial disentanglement for domain adaptation, ICML 2022*

Identifiability Theory

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

such that $\hat{z}_{s,j} = h_{s,i}(z_{s,i})$. For ease of exposition, we assume that the $\mathbf{z}_c$ and $\mathbf{z}_s$ correspond to components in $\mathbf{z}$ with indices $\{1, \ldots, n_c\}$ and $\{n_c + 1, \ldots, n\}$ respectively, that is, $\mathbf{z}_c = (z_i)_{i=1}^{n_c}$ and $\mathbf{z}_s = (z_i)_{i=n_c+1}^{n}$.

Theorem 4.1. *We follow the data generation process in Equation 1 and make the following assumptions:*

- *A1 (Smooth and Positive Density): The probability density function of latent variables is smooth and positive, i.e. $p_{\mathbf{z}|\mathbf{u}}$ is smooth and $p_{\mathbf{z}|\mathbf{u}} > 0$ over $\mathcal{Z}$ and $\mathcal{U}$.*
- *A2 (Conditional independence): Conditioned on $\mathbf{u}$, each z_i is independent of any other z_j for $i, j \in [n]$, $i \neq j$, i.e. $\log p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \sum_i^n q_i(z_i, \mathbf{u})$ where q_i is the log density of the conditional distribution, i.e., $q_i := \log p_{z_i|\mathbf{u}}$.*
- *A3 (Linear independence): For any $\mathbf{z}_s \in \mathcal{Z}_s \subseteq \mathbb{R}^{n_s}$, there exist $2n_s + 1$ values of $\mathbf{u}$, i.e., $\mathbf{u}_j$ with $j = 0, 1, \ldots, 2n_s$, such that the $2n_s$ vectors $\mathbf{w}(\mathbf{z}_s, \mathbf{u}_j) - \mathbf{w}(\mathbf{z}_s, \mathbf{u}_0)$ with $j = 1, ..., 2n_s$, are linearly independent, where vector $\mathbf{w}(\mathbf{z}_s, \mathbf{u})$ is defined as follows:*

$$\mathbf{w}(\mathbf{z}_s, \mathbf{u}) = \Big(\frac{\partial q_{n_c+1}(z_{n_c+1}, \mathbf{u})}{\partial z_{n_c+1}}, \ldots, \frac{\partial q_n(z_n, \mathbf{u})}{\partial z_n}, \frac{\partial^2 q_{n_c+1}(z_{n_c+1}, \mathbf{u})}{\partial z_{n_c+1}^2}, \ldots, \frac{\partial^2 q_n(z_n, \mathbf{u})}{\partial z_n^2} \Big). \tag{3}$$

By learning $(\hat{g}, p_{\hat{\mathbf{z}}_c}, p_{\hat{\mathbf{z}}_s|\mathbf{u}})$ to achieve Equation 2, $\mathbf{z}_s$ is component-wise identifiable.

Implementation with Modified VAE

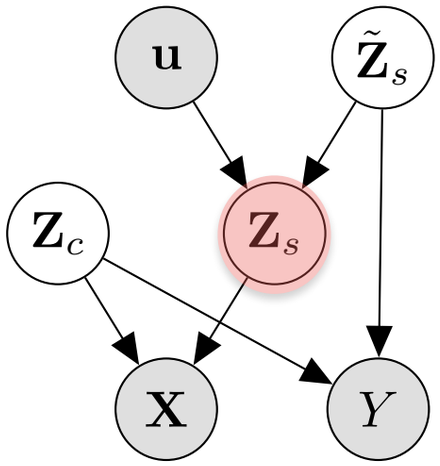
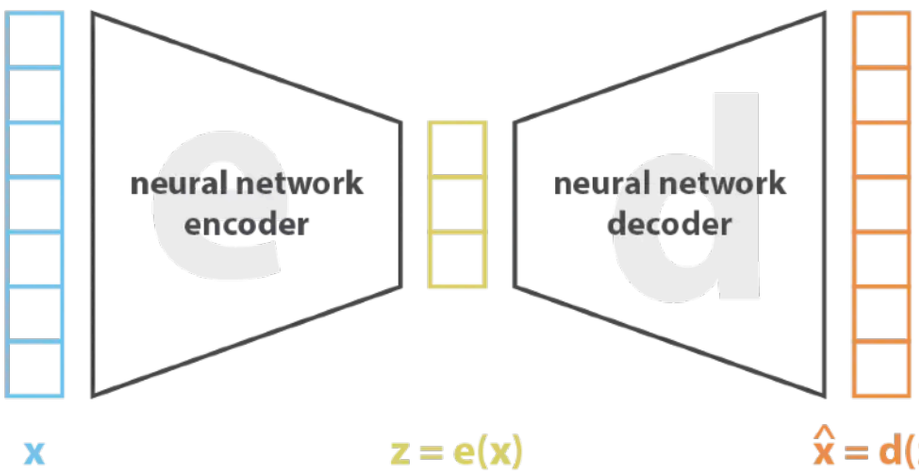


Figure 1. The generating process: The gray shade of nodes indicates that the variable is observable.



loss = $|| x - \hat{x} ||^2 = || x - d(z) ||^2 = || x - d(e(x)) ||^2$

Autoencoder

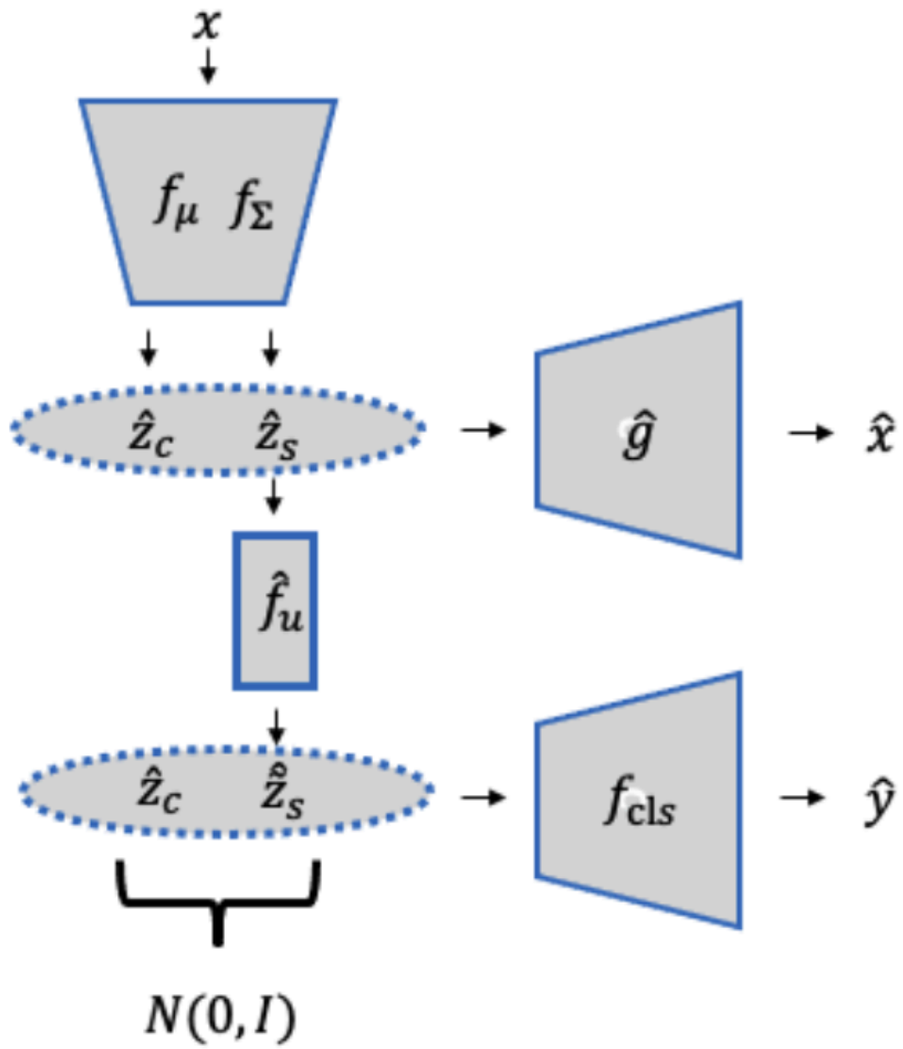
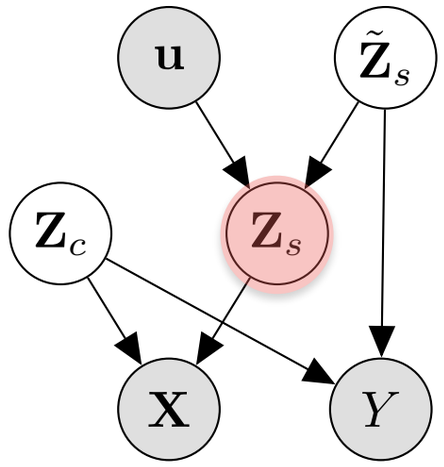


Figure 2. Diagram of our proposed method, **iMSDA**. We first apply the VAE encoder (f_μ, f_Σ) to encode $\mathbf{x}$ into $(\hat{\mathbf{z}}_c, \hat{\mathbf{z}}_s)$, which is further fed into the decoder $\hat{g}$ for reconstruction. In parallel, the changing part $\hat{\mathbf{z}}_s$ is passed through the flow model $f_\mathbf{u}$ to recover the high-level invariant variable $\hat{\hat{\mathbf{z}}}_s$. We use $(\hat{\mathbf{z}}_c, \hat{\hat{\mathbf{z}}}_s)$ for classification with the classifier f_{cls} and for matching $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with a KL loss.

Partial Disentanglement for Domain Adaptation

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



Models	→ Art	→ Clipart	→ Product	→ Realworld	Avg
Source Only (He et al., 2016)	64.58±0.68	52.32±0.63	77.63±0.23	80.70±0.81	68.81
DANN (Ganin et al., 2016)	64.26±0.59	58.01±1.55	76.44±0.47	78.80±0.49	69.38
DANN+BSP (Chen et al., 2019)	66.10±0.27	61.03±0.39	78.13±0.31	79.92±0.13	71.29
DAN (Long et al., 2015)	68.28±0.45	57.92±0.65	78.45±0.05	81.93±0.35	71.64
MCD (Saito et al., 2018)	67.84±0.38	59.91±0.55	79.21±0.61	80.93±0.18	71.97
M3SDA (Peng et al., 2019)	66.22±0.52	58.55±0.62	79.45±0.52	81.35±0.19	71.39
DCTN (Xu et al., 2018)	66.92±0.60	61.82±0.46	79.20±0.58	77.78±0.59	71.43
MIAN (Park & Lee, 2021)	69.39±0.50	63.05±0.61	79.62±0.16	80.44±0.24	73.12
MIAN-γ (Park & Lee, 2021)	69.88±0.35	**64.20±0.68**	80.87±0.37	81.49±0.24	74.11
iMSDA (Ours)	**75.77±0.21**	60.83±0.73	**84.13±0.09**	**84.83±0.12**	**76.39**

Table 2. Classification results on Office-Home. Backbone: Resnet-50. Baseline results are taken from (Park & Lee, 2021).

- *Xie, Kong, Gong, Zhang, "Multi-domain image generation and translation with identifiability guarantees", ICLR 2023*

A Weaker One: Subspace Identifiability for Domain Adaptation

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

Subspace Identification for Multi-Source Domain Adaptation

Zijian Li[2,3], **Ruichu Cai**[2,*], **Guangyi Chen**[3,1], **Boyang Sun**[3], **Zhifeng Hao**[4], **Kun Zhang**[3,1*]
[1] Carnegie Mellon University
[2] School of Computer Science, Guangdong University of Technology
[3] Mohamed bin Zayed University of Artificial Intelligence
[4] Shantou University

Abstract

Multi-source domain adaptation (MSDA) methods aim to transfer knowledge from multiple labeled source domains to an unlabeled target domain. Although current methods achieve target joint distribution identifiability by enforcing minimal changes across domains, they often necessitate stringent conditions, such as an adequate number of domains, monotonic transformation of latent variables, and invariant label distributions. These requirements are challenging to satisfy in real-world applications. To mitigate the need for these strict assumptions, we

Subspace Identifiability: Theory

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

$\hat{\mathbf{z}}_s$ and an invertible function $h_i : \mathbb{R}^n \to \mathbb{R}$, such that $z_{s,i} = h_i(\hat{\mathbf{z}}_s)$.

Theorem 1. ***(Subspace Identification of $\mathbf{z}_s$.)*** *We follow the data generation process in Figure 2 and make the following assumptions:*

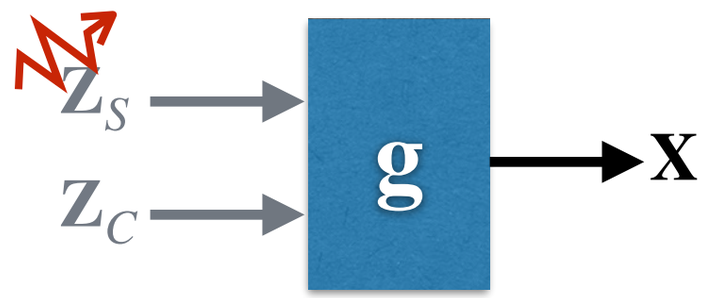
- *A1 (Smooth and Positive Density): The probability density function of latent variables is smooth and positive, i.e., $p_{\mathbf{z}|\mathbf{u}} > 0$ over $\mathcal{Z}$ and $\mathcal{U}$.*
- *A2 (Conditional independent): Conditioned on $\mathbf{u}$, each z_i is independent of any other z_j for $i, j \in \{1, \cdots, n\}, i \neq j$, i.e. $\log p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \sum_i^n q_i(z_i, \mathbf{u})$ where $q_i(z_i, \mathbf{u})$ is the log density of the conditional distribution, i.e., $q_i : \log p_{z_i|\mathbf{u}}$.*
- *A3 (Linear independence): For any $\mathbf{z}_s \in \mathcal{Z}_s \subseteq \mathbb{R}^{n_s}$, there exist $n_s + 1$ values of $\mathbf{u}$, i.e., $\mathbf{u}_j$ with $j = 0, 1, \cdots, n_s$, such that these n_s vectors $\mathbf{w}(\mathbf{z}, \mathbf{u}_j) - \mathbf{w}(\mathbf{z}, \mathbf{u}_0)$ with $j = 1, \cdots, n_s$ are linearly independent, where vector $\mathbf{w}(\mathbf{z}, \mathbf{u}_j)$ is defined as follows:*

$$\mathbf{w}(\mathbf{z}, \mathbf{u}) = \left(\frac{\partial q_1(z_1, \mathbf{u})}{\partial z_1}, \cdots, \frac{\partial q_i(z_i, \mathbf{u})}{\partial z_i}, \cdots \frac{\partial q_{n_s}(z_{n_s}, \mathbf{u})}{\partial z_{n_s}} \right), \tag{2}$$

By modeling the aforementioned data generation process, $\mathbf{z}_s$ is subspace identifiable.

Finding Changing Hidden Variables for Transfer Learning: *Minimal Change Principle*

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



- Changing components $\mathbf{Z}_S$ are identifiable; invariant part $\mathbf{Z}_C$ is identifiable up to its subspace
- Minimal Change Principle
  - What if we have **more** dimensions of $\mathbf{Z}_S$ than needed?
  - What if we have **fewer** dimensions of $\mathbf{Z}_S$ than needed?
  - Go with the right one!

Image Translation: How to Learn ‘Style’?

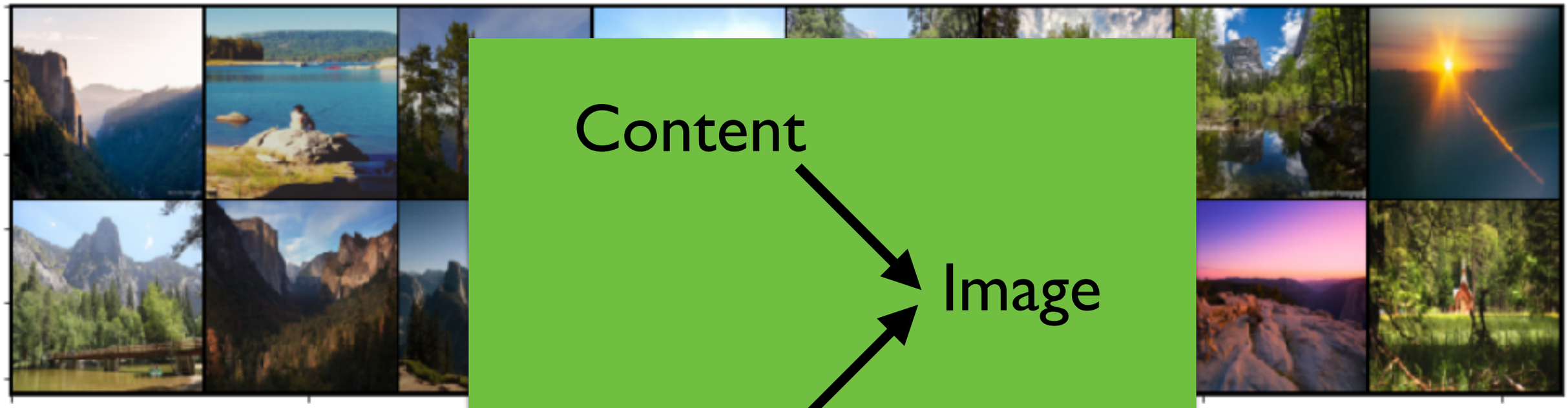
A simpler example:

Domain 1:

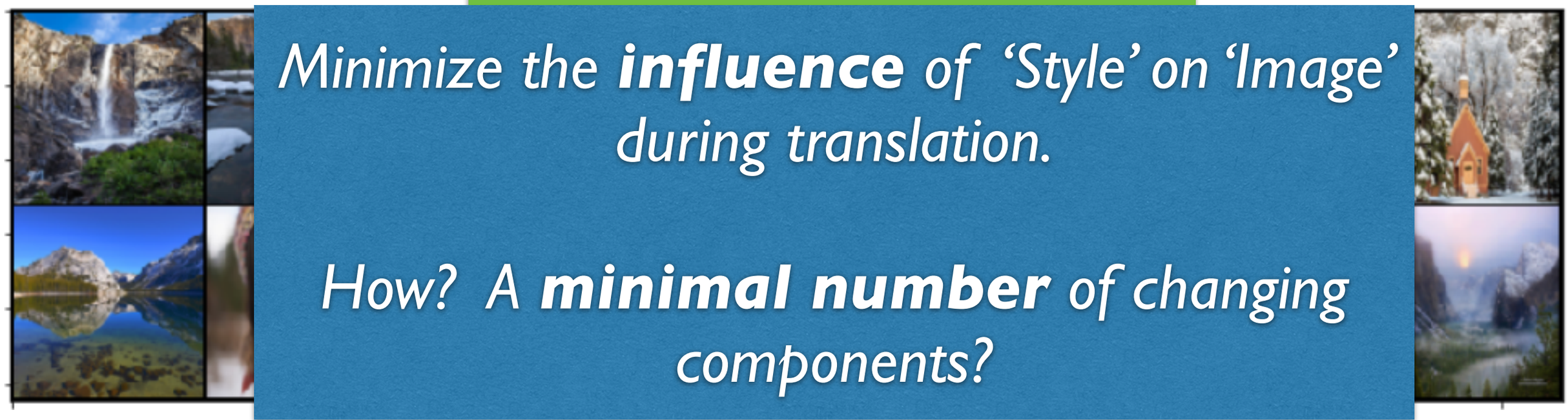
Domain 2:

Images from the winter season domain.

Minimal Changes Enables Identifiability



Im ain.



Content

Image

Style

*Minimize the **influence** of 'Style' on 'Image' during translation.*

*How? A **minimal number** of changing components?*

Images from the winter season domain.

MULTI-DOMAIN IMAGE GENERATION AND TRANSLATION WITH IDENTIFIABILITY GUARANTEES

Shaoan Xie[1], **Lingjing Kong**[1], **Mingming Gong**[3,2], and **Kun Zhang**[1,2]

[1] Carnegie Mellon University
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]The University of Melbourne
shaoan@cmu.edu, lingjingkong@cmu.edu,
mingming.gong@unimelb.edu.au, kunz1@cmu.edu

ABSTRACT

Multi-domain image generation and unpaired image-to-to-image translation are two important and related computer vision problems. The common technique for the two tasks is the learning of a joint distribution from multiple marginal distributions. However, it is well known that there can be infinitely many joint distributions that can derive the same marginals. Hence, it is necessary to formulate suitable constraints to address this highly ill-posed problem. Inspired by the recent advances in nonlinear Independent Component Analysis (ICA) theory, we propose a new method to learn the joint distribution from the marginals by enforcing a specific type of minimal change across domains. We report one of the first results connecting multi-domain generative models to identifiability and shows

Sample Images Generated by Generative Adversarial Networks (GANs)

Images generated by a [GAN created by NVIDIA](#).

Generative Adversarial Networks (GANs)

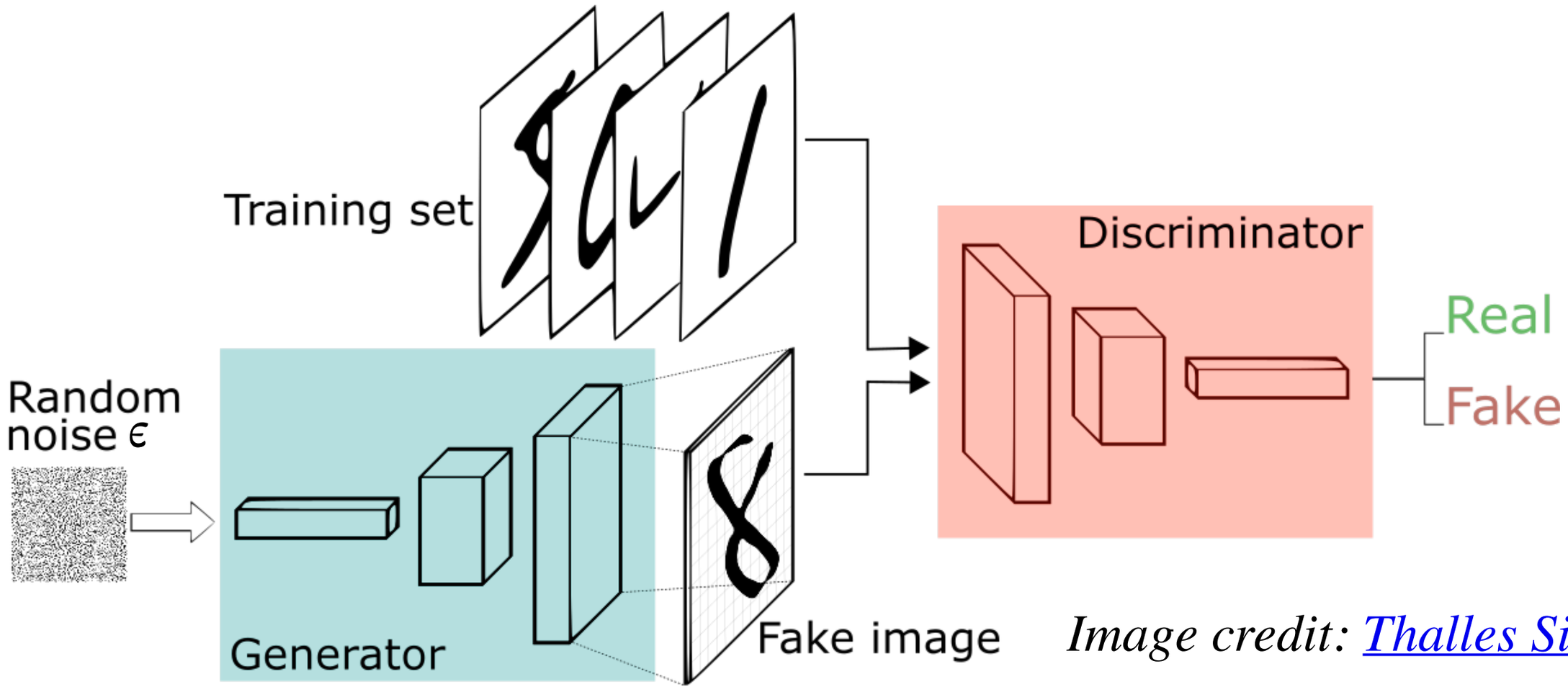
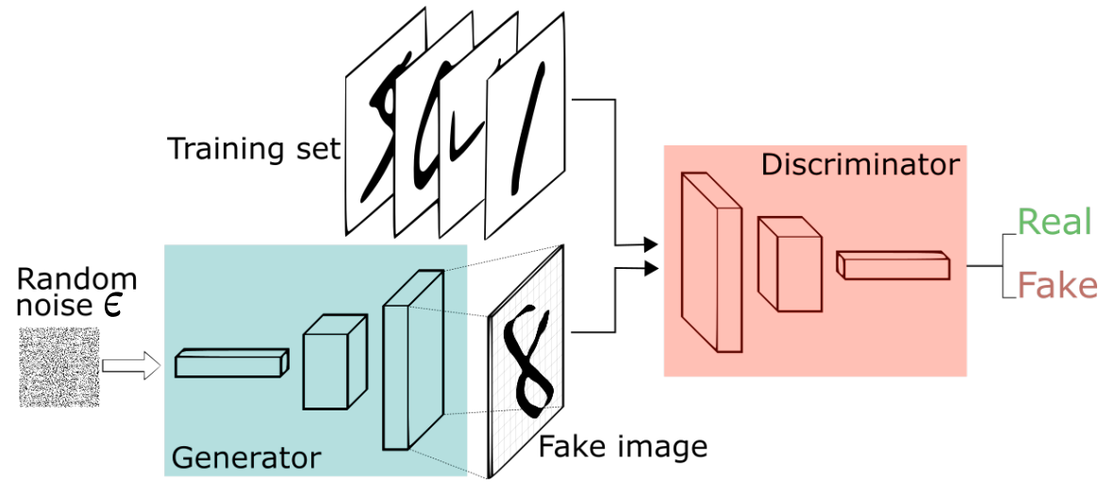


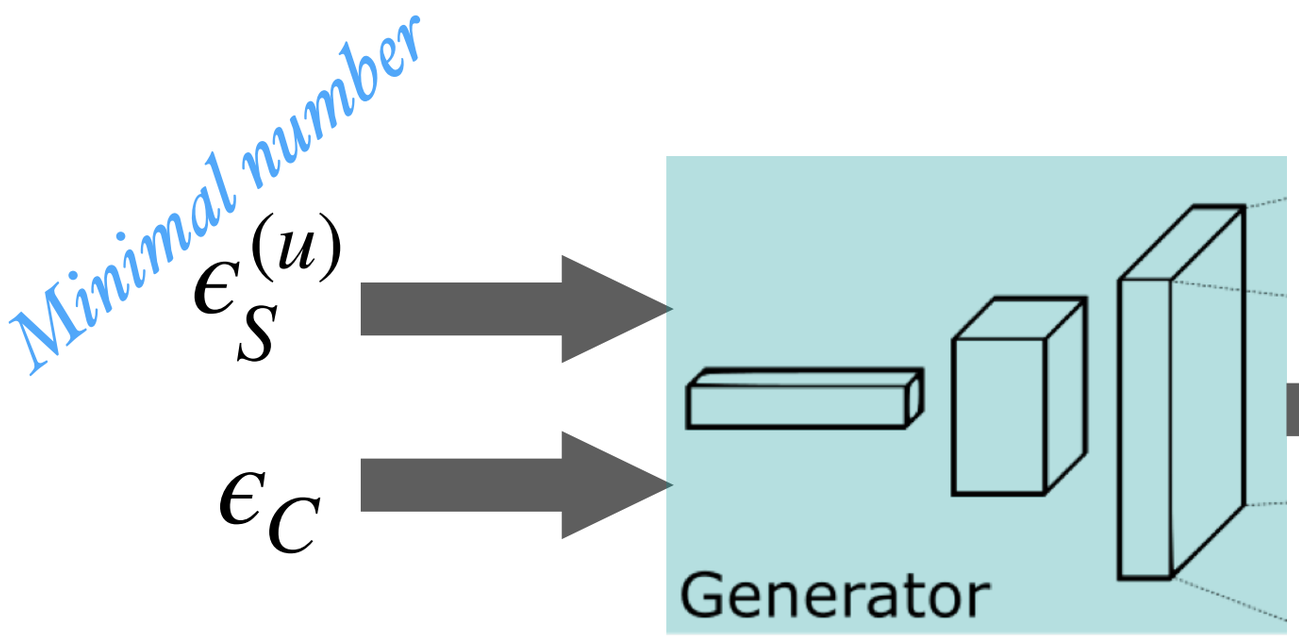
Image credit: Thalles Silva

Minimax game which G wants to minimize V while D wants to maximize it:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

GAN-Based Implementations





- Match the data distribution across domains, while the dimensionality of $\epsilon_S^{(u)}$ is as small as possible (minimal changes across domains **controlled by λ;** no penalty when λ=0)

- Correspondence relations among domains are identifiable

Multi-domain Image Generation & Translation with Identifiability Guarantees

- Idea: Matching the distributions across domains with a minimal number of changing components
- Correspondence info (joint distribution) identifiable under mild assumptions
- Example: Generating female & males images with the same "content"

Ours (λ=0.1)

StyleGAN2-ADA

TGAN

- *Xie, Kong, Gong, Zhang, "Multi-domain image generation and translation with identifiability guarantees", ICLR 2023*

More results…

Figure 10: CelebA-HQ. Without the sparsity regularization, i.e., $\lambda = 0$, we observe some unnecessary changes between the image tuples in each row. For example, e.g.,the added sun-glasses and skin color change in the first row. TGAN changes the background (first row of third panel). CoGAN changes the skin color (second row, second panel).

More results…

Alternative strategy: learning hidden representations from IID data.

Talk to Yujia about it. ;)

Figure 11: AFHQ. StyleGAN2-ADA changes animal poses in many examples, e.g., second and third row of first panel. Our base ($\lambda = 0$) also changes the poses, e.g., first and third row of second panel. CoGAN and TGAN are slightly better in preserving poses but we can observe that some generated images are unrealistic. For example, the wolf (first row, third panel of TGAN) and the dog (third row, third panel of CoGAN).

CRL from Changes: Outline

I., but non-I.D.	No	No	More informative than MEC (CD-NOD)
	Yes		May have unique identifiability
	No	Yes	Changing subspace identifiable
	Yes		Variables in changing relations identifiable

- CD-NOD (Causal Discovery from Nonstationary/Heterogenuous data)
- Nonlinear ICA with partial changes across domains
  - Partial disentanglement
  - Domain adaptation, image translation, and multi-domain data generation
  - Learning from text-image pairs
- A general setting
- Connection to the IID case: Synergy between minimal changes and sparsity

Motivation: Controllability for Image Generation / Editing

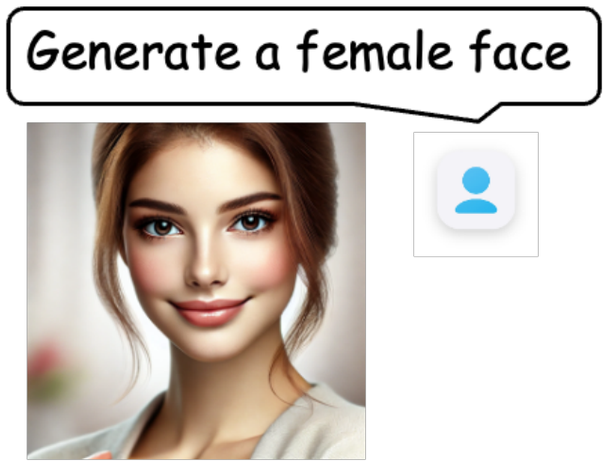
- Existing text-to-image (T2I) models are not controllable: editing a specific feature through text often causes unwanted changes
- Example:

“Happy”

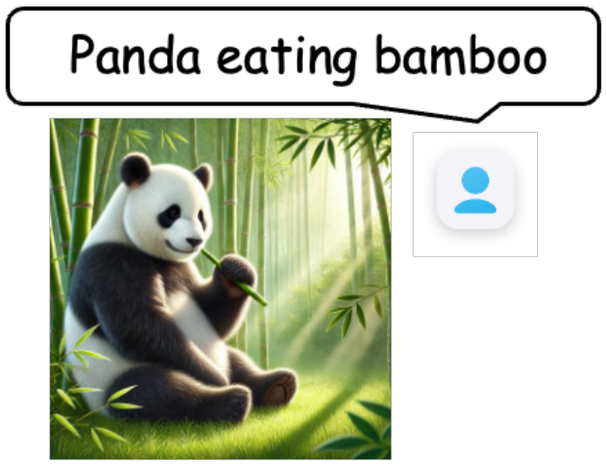
“Angry”

“Surprised”

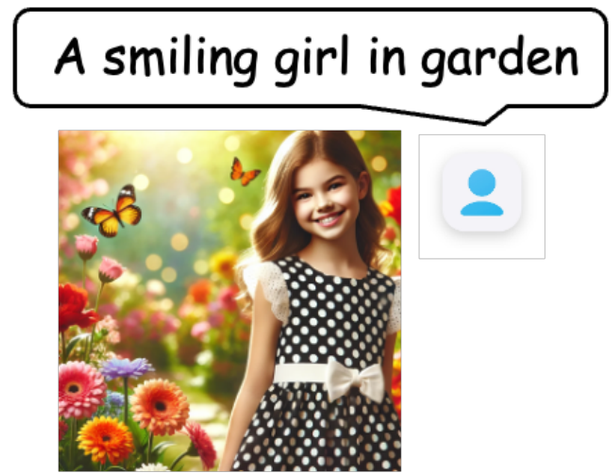
Prevailing generative AI tools: not controllable



Example1: Add Mustache



Example2: Change the style

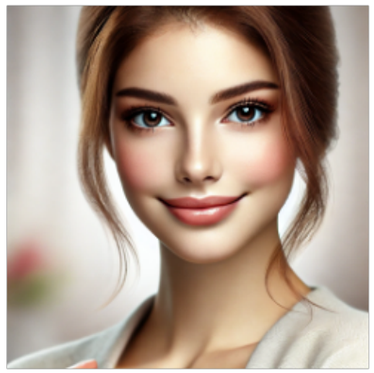


Example3: Change facial expression

Our Causal GenAI Enables Precise Control & Refinement

Example1: Add Mustache

Generate a female face



OpenAI

Can you add mustache ?

Example 3: Change facial expression

A smiling girl in garden

OpenAI

She looks surprised

Our Causal GenAI Enables Precise Control & Refinement

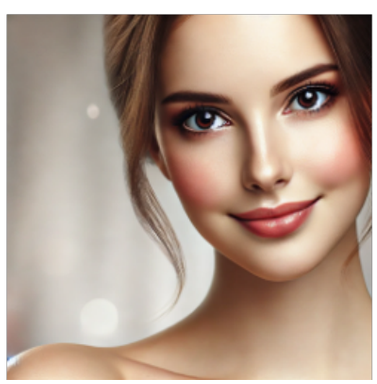
Example1: Add Mustache

Generate a female face

OpenAI

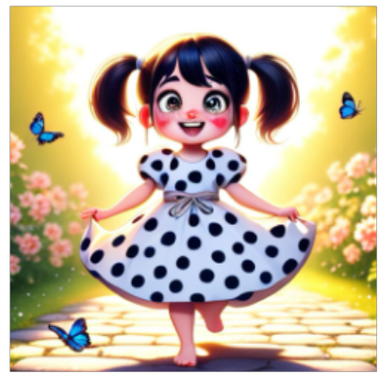
Can you add mustache ?

Ours



Example 3: Change facial expression

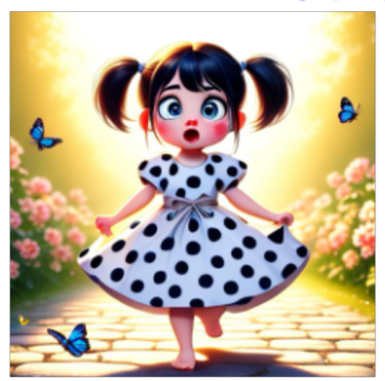
A smiling girl in garden



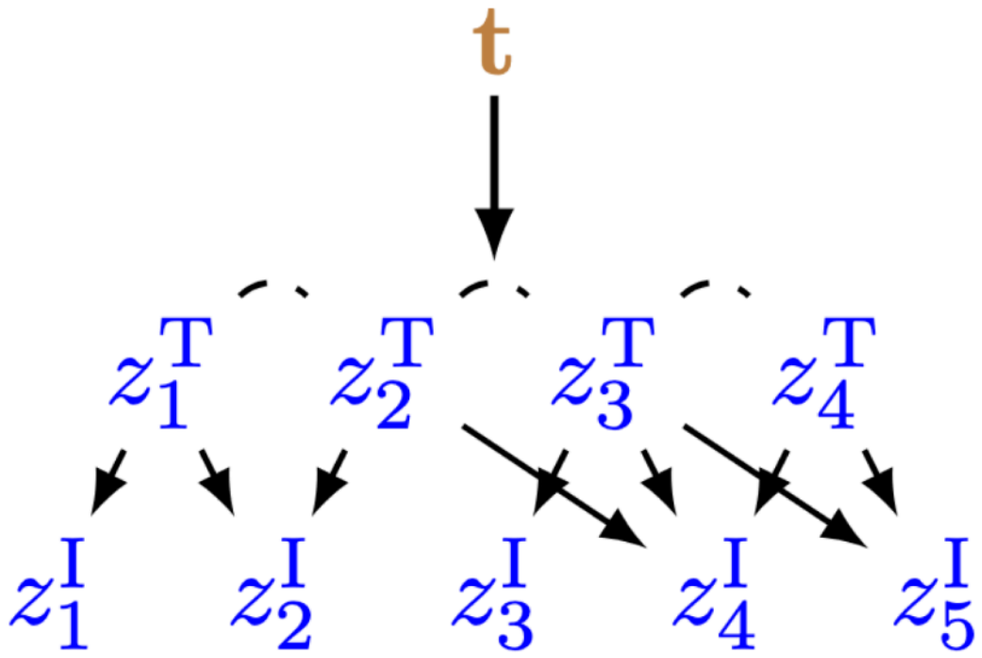
OpenAI

She looks surprised

Ours



From Text to Images: The Process



t: text

z_i^{T}: atomic textual concepts

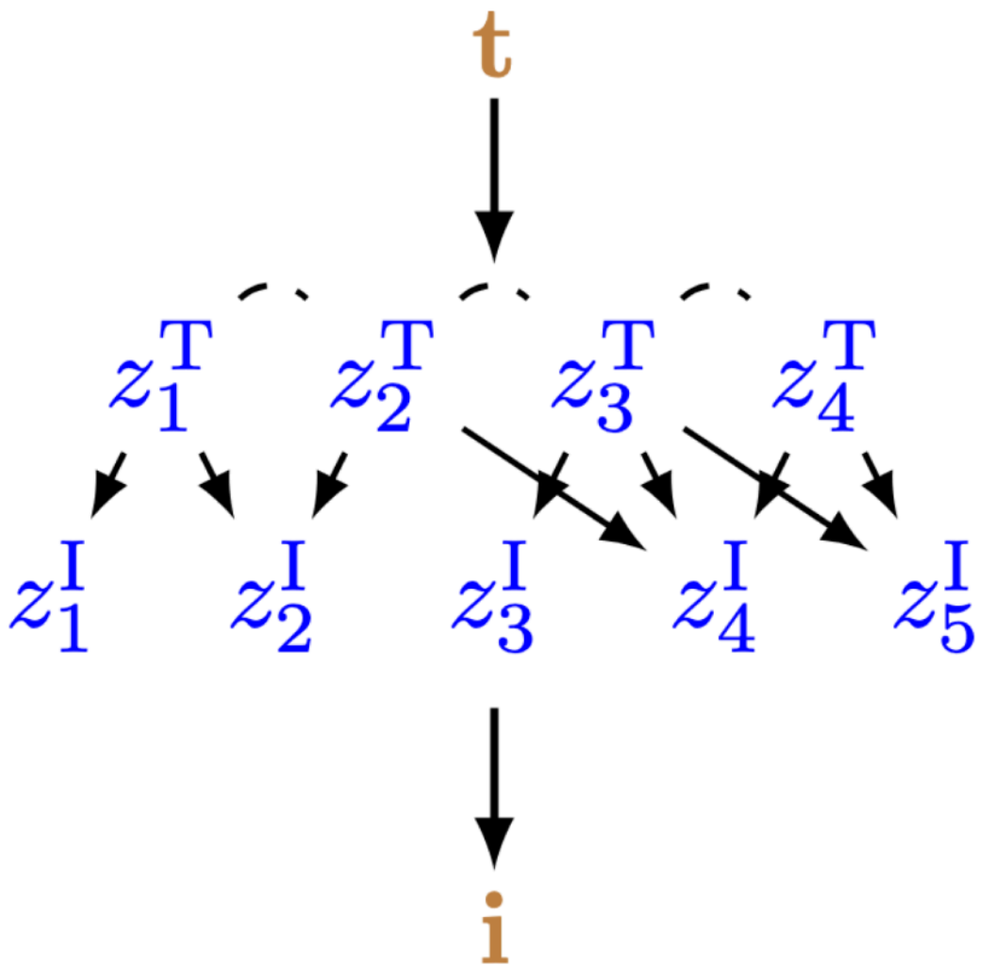
z_j^{I}: atomic visual concepts

i: images

- Text and images have atomic concepts
- Textual atomic concepts determine their visual counterparts: why?

– *Xie, Kong, Zheng, Tang, Xing, Chen, and Zhang, under submission*

Learning Identifiable Concepts for Controllability



t: text

z_i^{T}: atomic textual concepts

z_j^{I}: atomic visual concepts

i: images

Certain sparsity constraints on the cross links + conditional independence of image concepts $\Rightarrow$ **identifiable** concepts:

1. Learning *disentangled*, *atomic* concepts z_m^{T} and z_n^{I}.

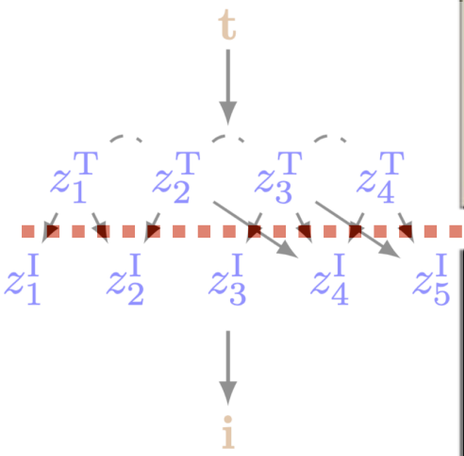
2. *Aligning* them.

Results: Controllable Generation

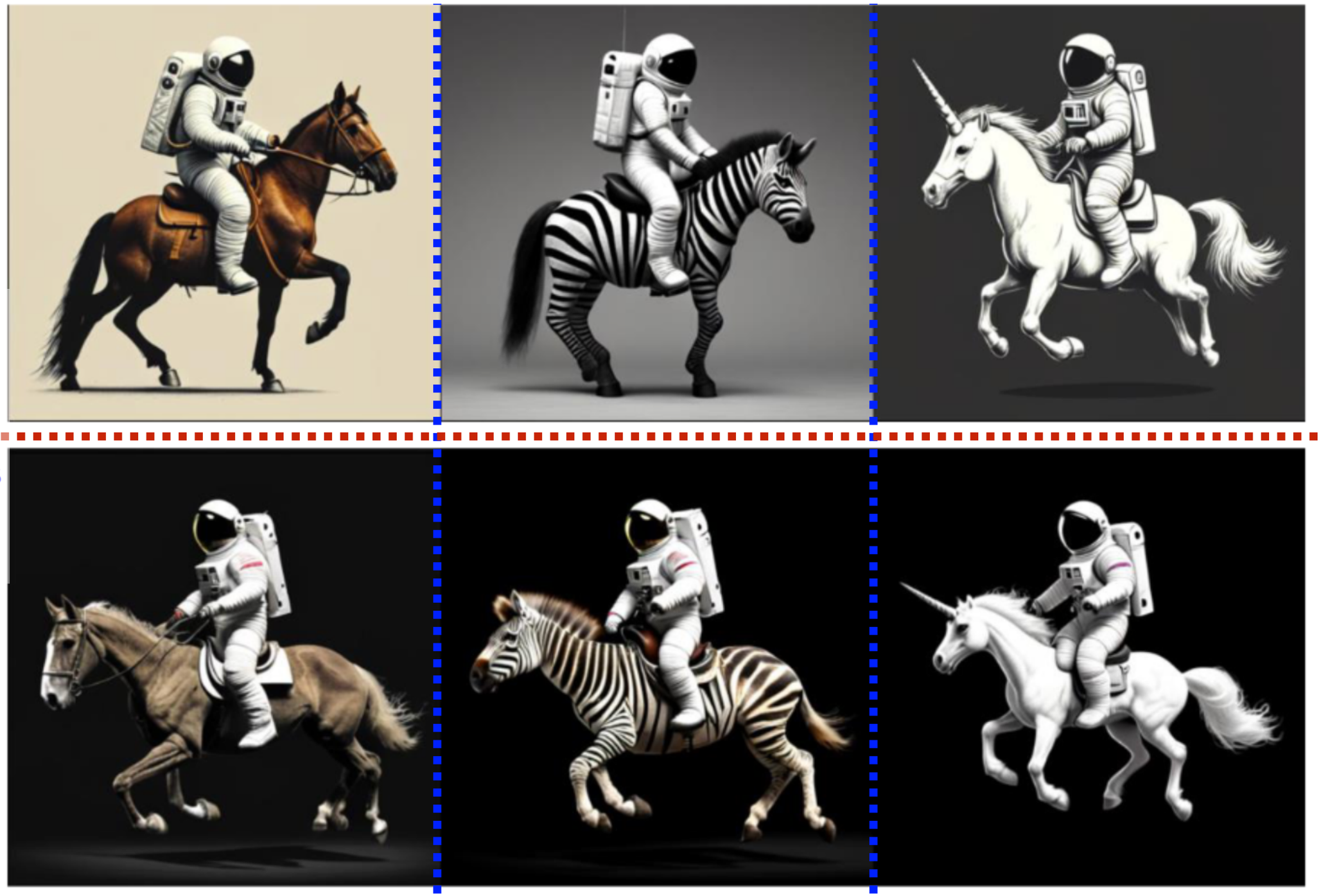
Our method can make only necessary changes without affecting other attributes!

Illustration: Importance of Sparsity

Without sparse interaction



With sparse interaction



"An astronaut is riding a horse."

"An astronaut is riding a zebra"

"An astronaut is riding a unicorn."

Remember these results?

By Stable Diffusion: One Year Ago

- Prompt: a peacock eating ice cream

By DALL·E 3: Three Months Ago

- Prompt: a peacock eating ice cream

Let Peacock Eat Ice Cream, Controllably

- Prompt: a peacock eating ice cream
- Prompt: a peacock eating **white** ice cream

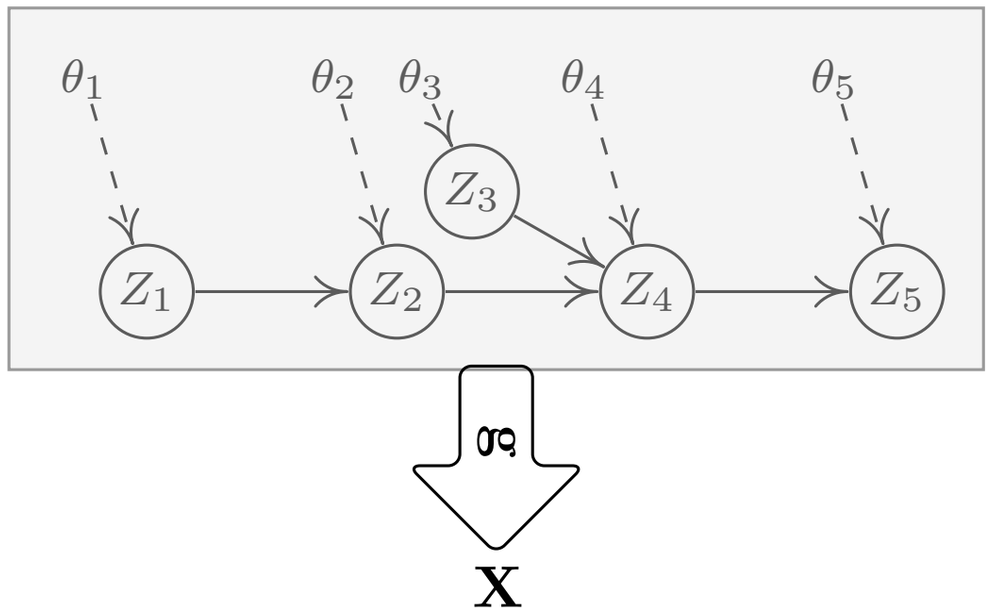
CRL from Changes: Outline

I., but non-I.D.	No	No	More informative than MEC (CD-NOD)
	Yes		May have unique identifiability
	No	Yes	Changing subspace identifiable
	Yes		Variables in changing relations identifiable

- CD-NOD (Causal Discovery from Nonstationary/Heterogenuous data)
- Nonlinear ICA with partial changes across domains
  - Partial disentanglement
  - Domain adaptation, image translation, and multi-domain data generation
  - Learning from text-image pairs
- A general setting
- Connection to the IID case: Synergy between changes and sparsity

Causal Representation Learning from Multiple Distributions: A *General* Setting

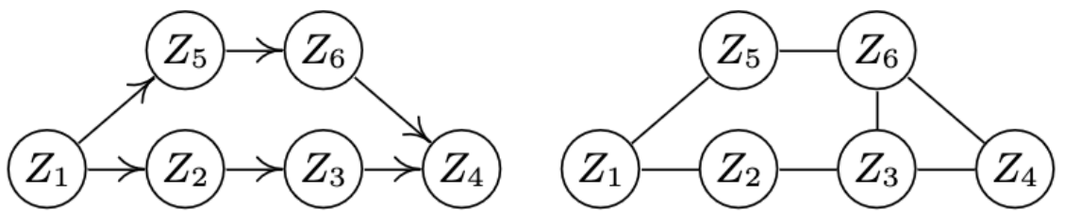
i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



- Goal: Uncovering hidden variables Z_i with changing causal relations from $\mathbf{X}$ in nonparametric settings

- What is identifiable?
  - Markov network of Z_i
  - Each estimated variable $\tilde{Z}_i$ is a function of Z_i and it intimate neighbors

- In this example, each Z_i ($i \neq 4$) can be recovered up to component-wise transformation



(a) $\mathcal{G}_Z$, the DAG over true latent variables Z_i.

(b) The corresponding Markov network $\mathcal{M}_Z$.

- *Zhang, Xie, Ng, Zheng, "Causal Representation Learning from Multiple Distributions: A General Setting," ICML 2024*

CRL from Changes: Outline

I., but non-I.D.	No	No	More informative than MEC (CD-NOD)
	Yes		May have unique identifiability
	No	Yes	Changing subspace identifiable
	Yes		Variables in changing relations identifiable

- CD-NOD (Causal Discovery from Nonstationary/Heterogenuous data)
- Nonlinear ICA with partial changes across domains
  - Partial disentanglement
  - Domain adaptation, image translation, and multi-domain data generation
  - Learning from text-image pairs
- A general setting
- Connection to the IID case: Synergy between minimal changes and sparsity

SYNERGY BETWEEN SUFFICIENT CHANGES AND SPARSE MIXING PROCEDURE FOR DISENTANGLED REPRESENTATION LEARNING

Zijian Li[†•*] **Shunxing Fan**[•*] **Yujia Zheng**[†] **Ignavier Ng**[†] **Shaoan Xie**[†] **Guangyi Chen**[†•]
Xinshuai Dong[†] **Ruichu Cai**[‡] **Kun Zhang**[†•]
[†]Carnegie Mellon University, Pittsburgh PA, USA
[•]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[‡]Guangdong University of Technology, Guangzhou, China

ABSTRACT

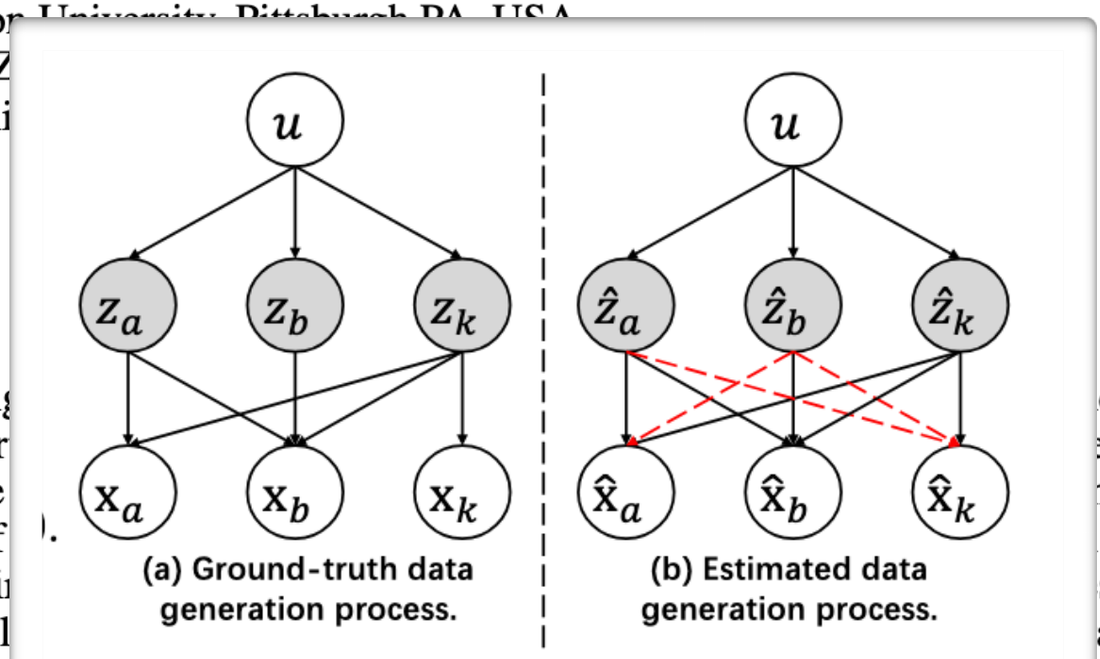
Disentangled representation learning aims to uncover latent variables underlying the observed data, and generally speaking, rather strong assumptions are needed

Is it possible to leverage both the distribution changeability & sparsity constraint in a complementary, principled way to learn disentangled representations with identifiability guarantees?

liability. Specifically, when conditioned on auxiliary variables, the sparse mixing procedure assumption provides structural constraints on the mapping from estimated to true latent variables and hence compensates for potentially insufficient distribution changes. Building on this insight, we propose an identifiability theory with less restrictive constraints regarding distribution changes and the sparse mixing procedure, enhancing applicability to real-world scenarios. Additionally, we

SYNERGY BETWEEN SUFFICIENT CHANGES AND SPARSE MIXING PROCEDURE FOR DISENTANGLED REPRESENTATION LEARNING

Zijian Li[†•*] **Shunxing Fan**[•*] **Yujia Zheng**[†] **Ignavier Ng**[†] **Shaoan Xie**[†] **Guangyi Chen**[†•]
Xinshuai Dong[†] **Ruichu Cai**[‡] **Kun Zhang**[†•]
[†]Carnegie Mellon University, Pittsburgh PA, USA
[•]Mohamed bin Z
[‡]Guangdong Uni



(a) Ground-truth data generation process.

(b) Estimated data generation process.

Figure 1: Example for Theorem 1 with ground-truth and estimated data generation processes. The red dashed lines denote the redundant estimated mixing edges.

Disentang derlying
the obser e needed
to ensure he distri-
bution of indices,
but acqui s exploit
structural aints are
usually (p o seem-
ingly unre e identi-
fiability. e mixing
procedure om esti-
mated to ufficient
distributi y theory
with less restrictive constraints regarding distribution changes and the sparse mixing procedure, enhancing applicability to real-world scenarios. Additionally, we

Summary

- Remember Semmelweis?
  - Learning hidden causal factor from different distributions
- How can CRL benefit from distribution shift?
  - Constraints on the changes!
    - E.g., only $p(S_i)$ change
  - Minimal changes for "concept" identifiability
  - Learning atomic textual and visual concepts and connections
- Unification: the benefit from **sparse** mixing procedure & **minimal** changes