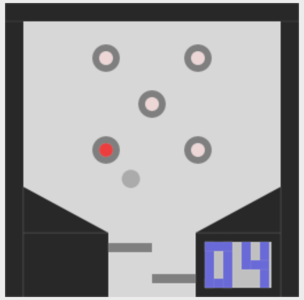


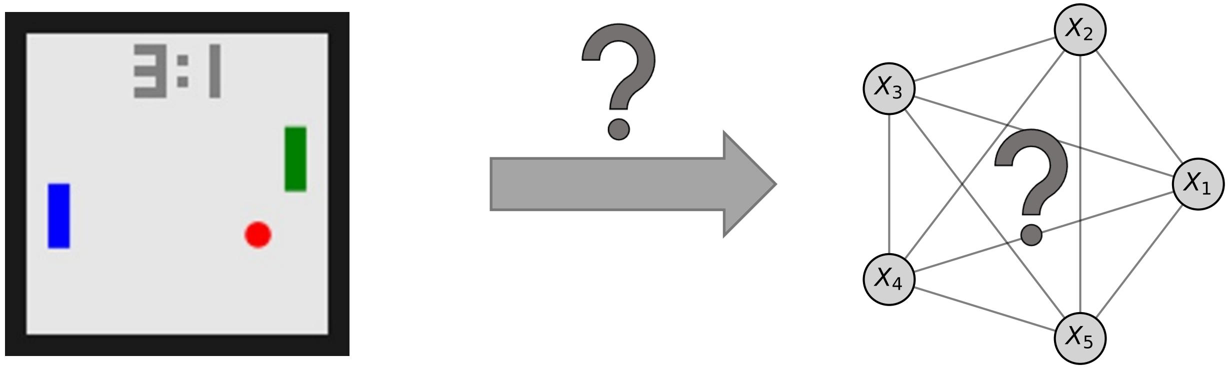
CITRIS: Causal Identifiability from Temporal Intervened Sequences

Phillip Lippe

06. October 2022

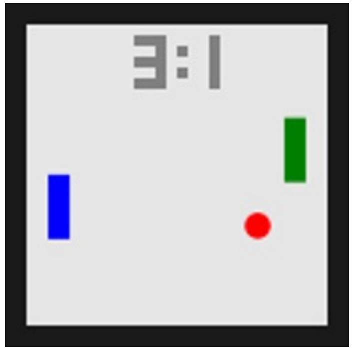
Causal Representation Learning

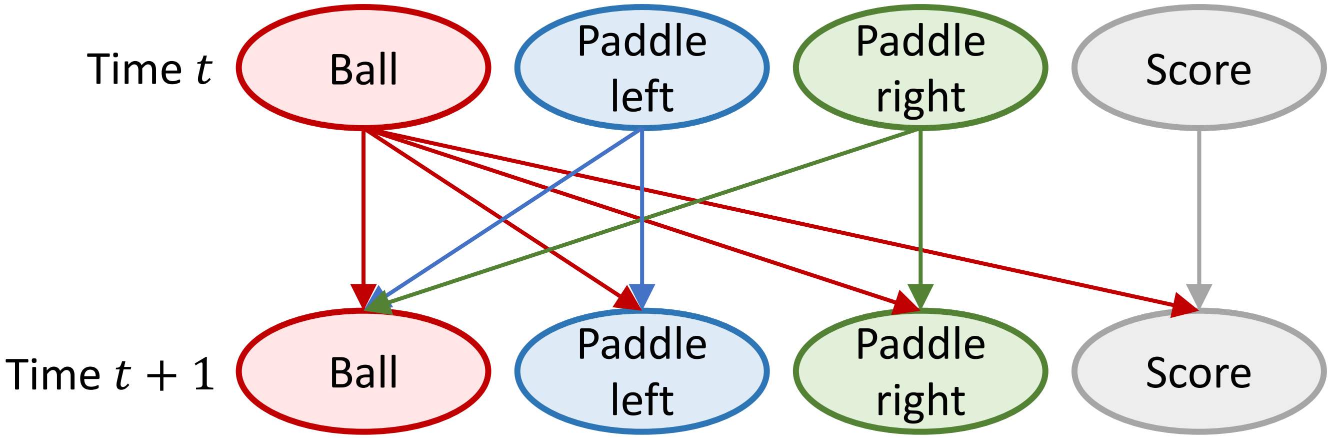
- Given high-dimensional observations of a (dynamical) system, what is its latent causal structure?
- Crucial for reasoning, planning, generalization



Causal Representation Learning

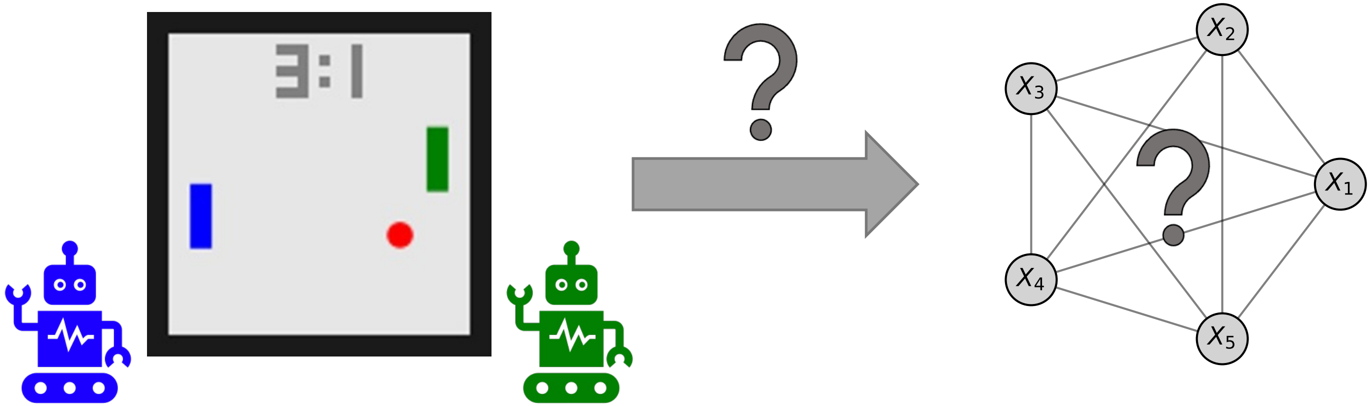
- Given high-dimensional observations of a (dynamical) system, what is its latent causal structure?
- Crucial for reasoning, planning, generalization, identifying cause-effect relations, etc.





Causal Representation Learning
Challenges

- High-dimensional input ↔ low-dimensional causal system
- Causal variables depend on each other
- Multiple (non-)causal representations can describe the same system
- Is a ‘causal’ representation unique?



Causal Representation Learning

Forms

Counterfactual CRL

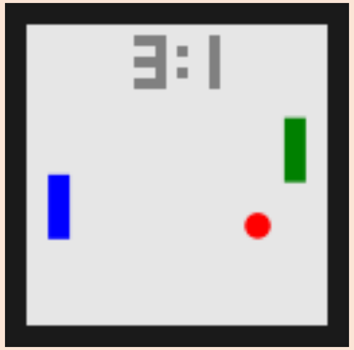
- Pairs of images where only a subset of variables change
- Requires a lot of control over system; not possible in real world (Pearl, 2009)

Examples: [Brehmer et al., 2022; Locatello et al., 2020; von Kügelgen et al., 2021; Ahuja et al., 2022]

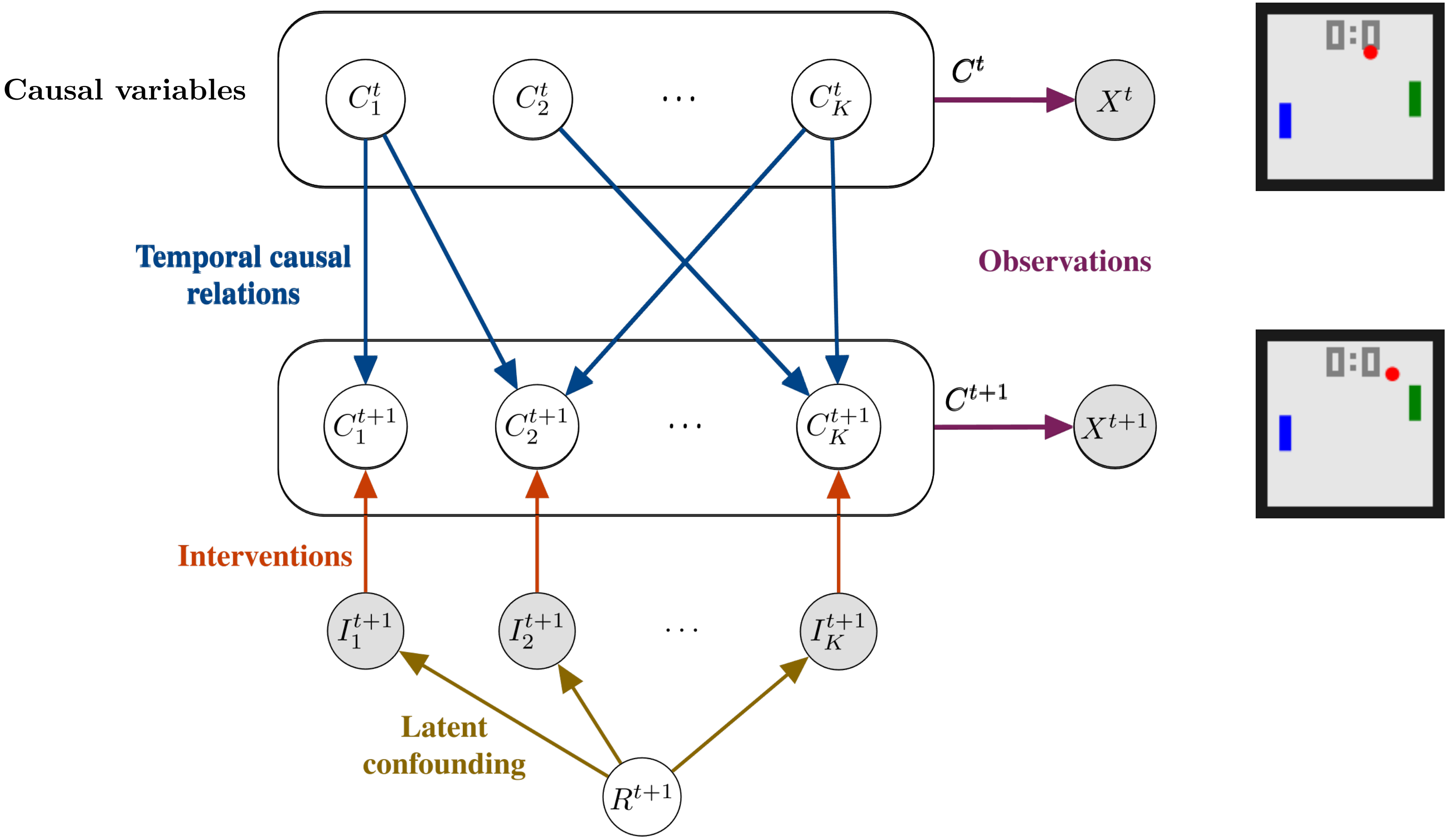
Temporal CRL

- Temporal sequences; all causal variables evolve over time
- Common RL environments
- Temporality gives strong bias

Examples: [Lippe et al., 2022ab; Lachapelle et al., 2022 ab; Yao et al., 2022ab; Khemakhem et al., 2020; Hyvärinen et al.; 2019]



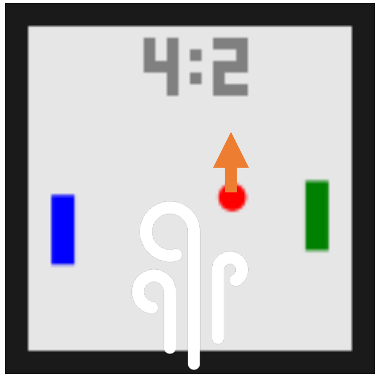
Causal Identifiability from Temporal Intervened Sequences
Setup

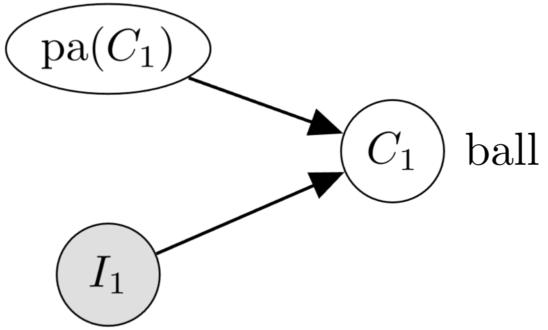


Causal Identifiability from Temporal Intervened Sequences

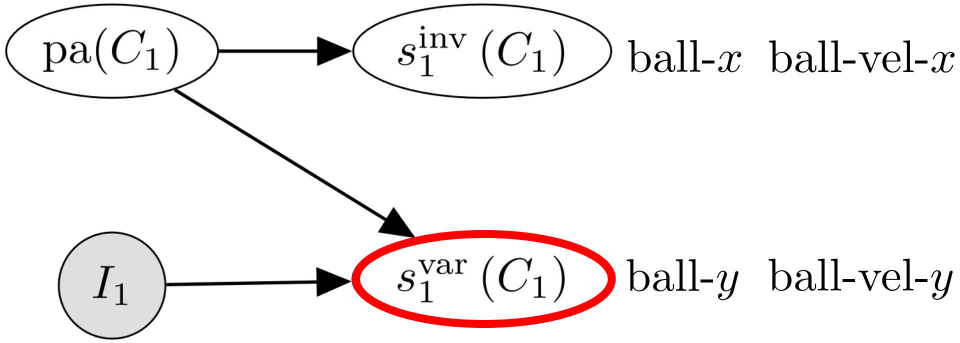
Minimal Causal Variables

- Abstraction ⇒ Multidimensional causal variables
- Identifying abstraction level ⇒ Interventions
- Augment causal graph with intervention targets
  - $I_1 = 1$ ⇒ Intervention on C_1
  - $I_1 = 0$ ⇒ Passively observing C_1
- Minimal causal variable $s_1^{\text{var}}(C_1)$: intervention-dependent part of a multidimensional causal variable





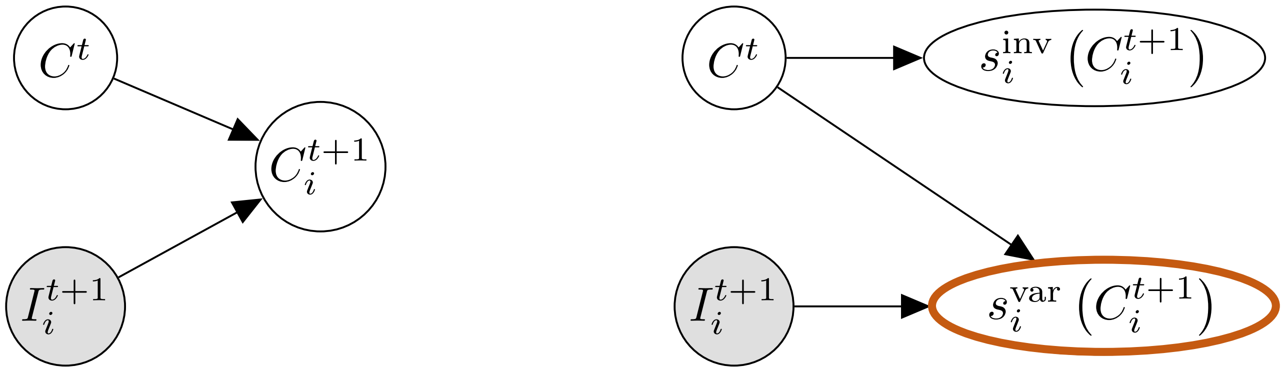
(a) Original causal graph of C_1



(b) Minimal causal split graph of C_1

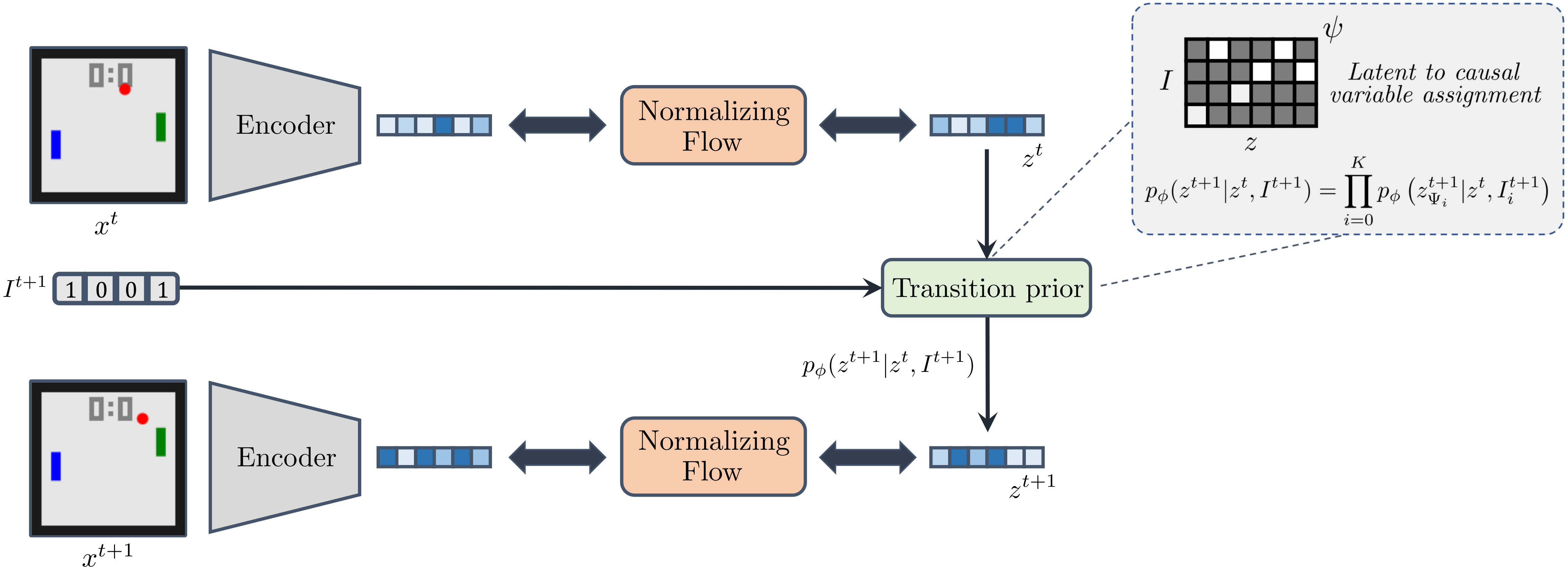
Causal Identifiability from Temporal Intervened Sequences
Theoretical Results

- Main theoretical result: we can identify the ***minimal causal variables*** up to invertible, component-wise transformations if:
  - No intervention target I_i^{t+1} is a deterministic function of any other:
  $$C_i^{t+1} \not\perp\!\!\!\perp I_i^{t+1} | C^t, I_j^{t+1}$$
  - Following intervention design, $\lfloor \log_2 K \rfloor + 2$ experiments are sufficient for this [Lippe et al., 2022c]

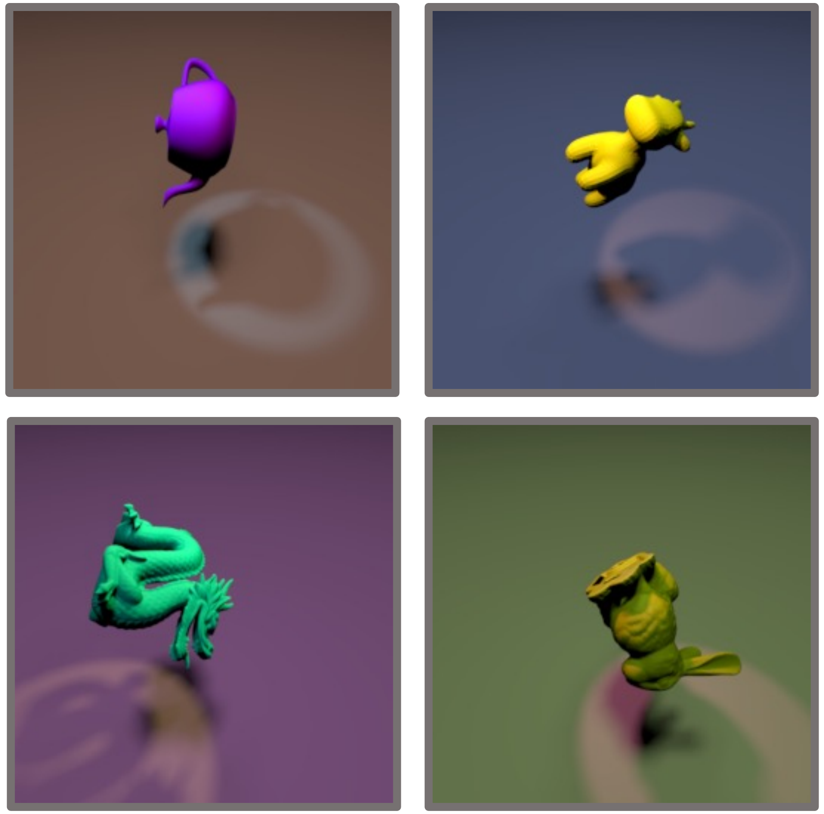


(a) Original causal graph of C_i (b) Minimal causal split graph of C_i

CITRIS Architecture
CITRIS-NF



CITRIS Experiments
Temporal Causal3DIdent



Causal Factors

object-shape
object-position
object-hue
object-rotation
spotlight-hue
spotlight-rot
background-hue

categorical
continuous
angle / circular

Zimmermann, Roland S., et al. "Contrastive learning inverts the data generating process." *ICML*, 2021.

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

CITRIS Experiments
Temporal Causal3DIdent

Novel combinations of causal factors

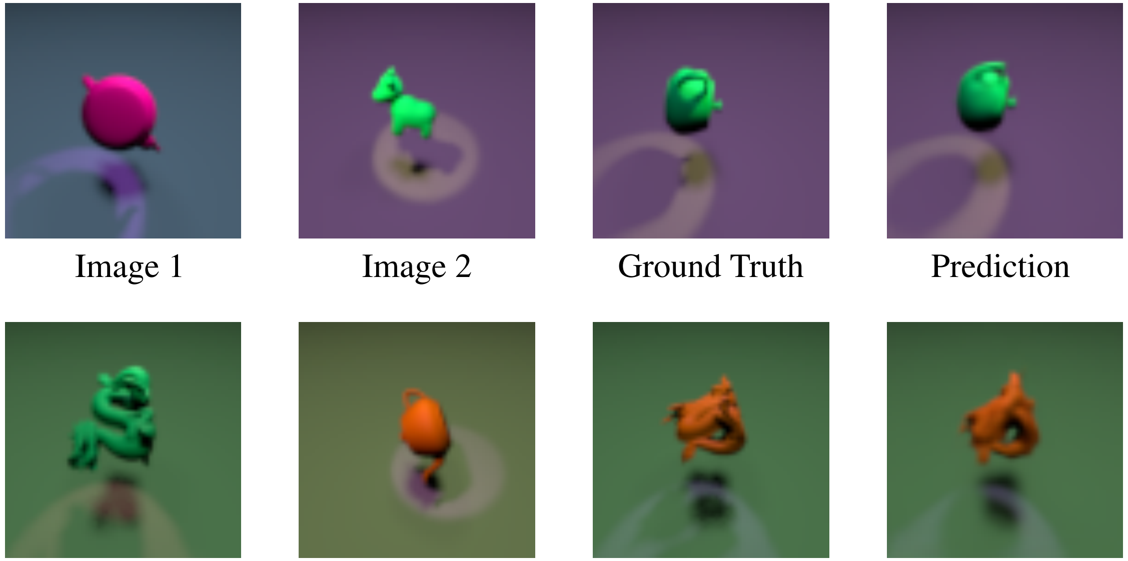
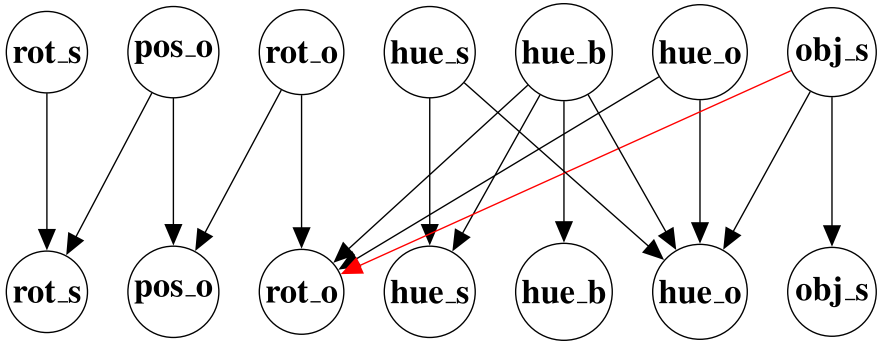


Image 1 | Image 2 | Ground Truth | Prediction

Image 1 | Image 2 | Ground Truth | Prediction

Learned Causal Graph



Summary

- **CITRIS**: Identify multidimensional causal variables from temporal sequences with soft interventions and known intervention targets [Lippe et al., 2022a]
- Identifies minimal causal variables, i.e., part of the variables that depends on interventions
- CITRIS-NF scales to visually complex scenes with pretrained autoencoder

- CITRIS provides flexible, extendable framework
  - iCITRIS: Extension to instantaneous effects within a time step [Lippe et al., 2022b]
  - Intervention Design for finding most efficient experiment set [Lippe et al., 2022c]

References

[Lippe et al., 2022a] Lippe, Phillip, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. "**CITRIS: Causal Identifiability from Temporal Intervened Sequences**." In International Conference on Machine Learning, pp. 13557-13603. PMLR, 2022.

[Lippe et al., 2022b] Lippe, Phillip, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. "**iCITRIS: Causal Representation Learning for Instantaneous Temporal Effects.**" First Workshop on Causal Representation Learning (CRL), UAI 2022.

[Lippe et al., 2022c] Lippe, Phillip, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. ”**Intervention Design for Causal Representation Learning.**" First Workshop on Causal Representation Learning (CRL), UAI 2022.

[Brehmer et al., 2022] Brehmer, Johann, Pim de Haan, Phillip Lippe, Taco Cohen. “**Weakly supervised causal representation learning**.“ Advances in Neural Information Processing Systems, NeurIPS 2022.