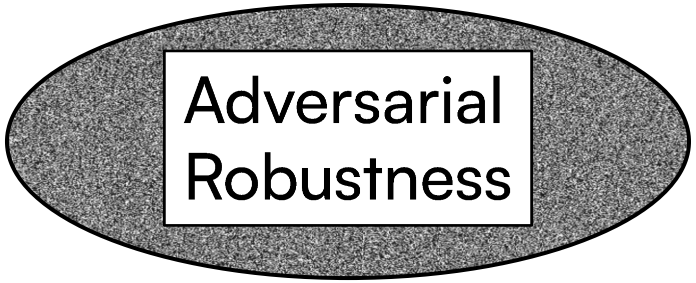
Provable Compositional Generalization for
Object-Centric Learning

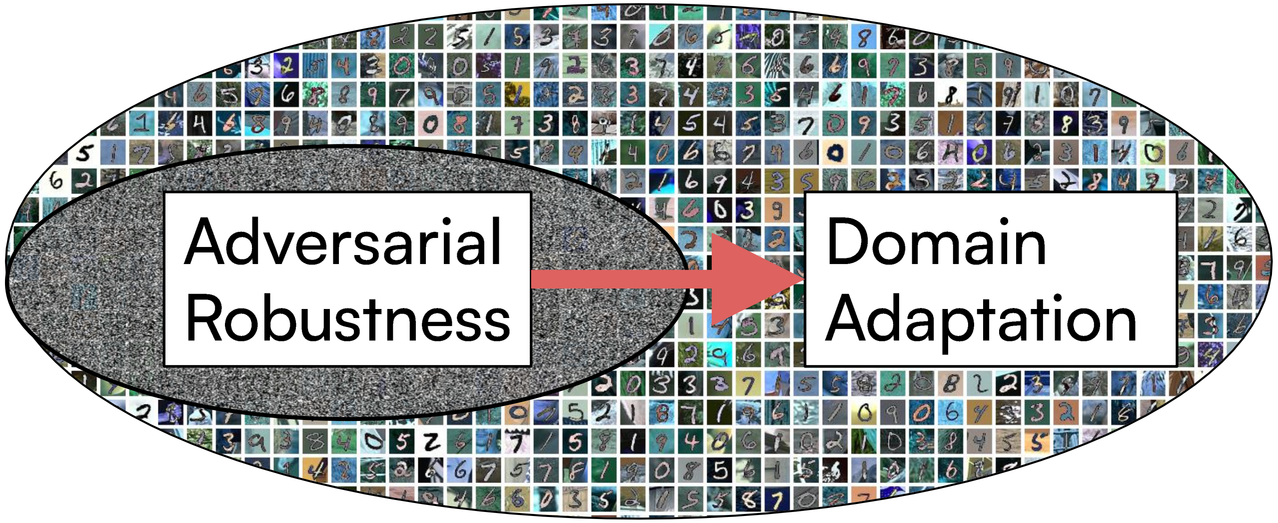
Talk by Alexander (Sasha) Panfilov
05/11/2024, EPFL, Lausanne

Disclaimer

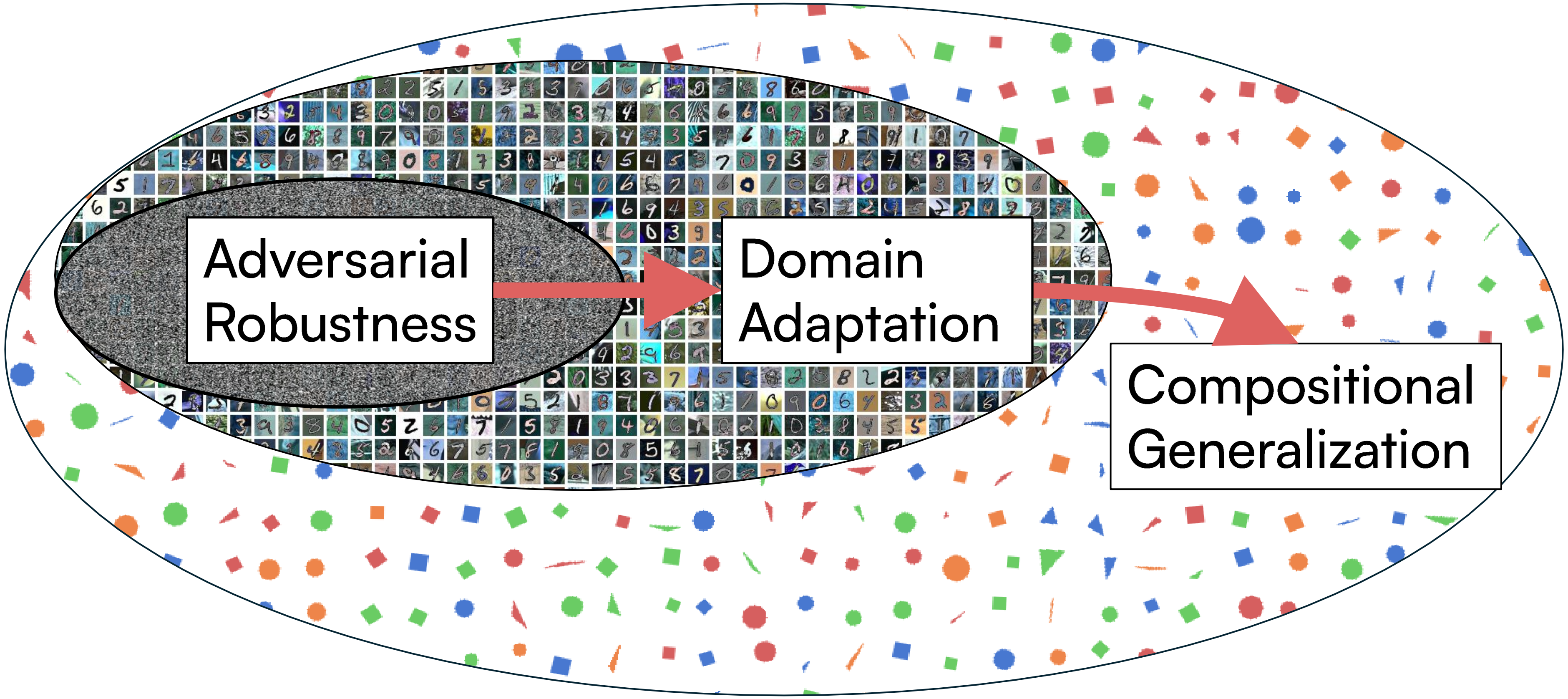
Disclaimer



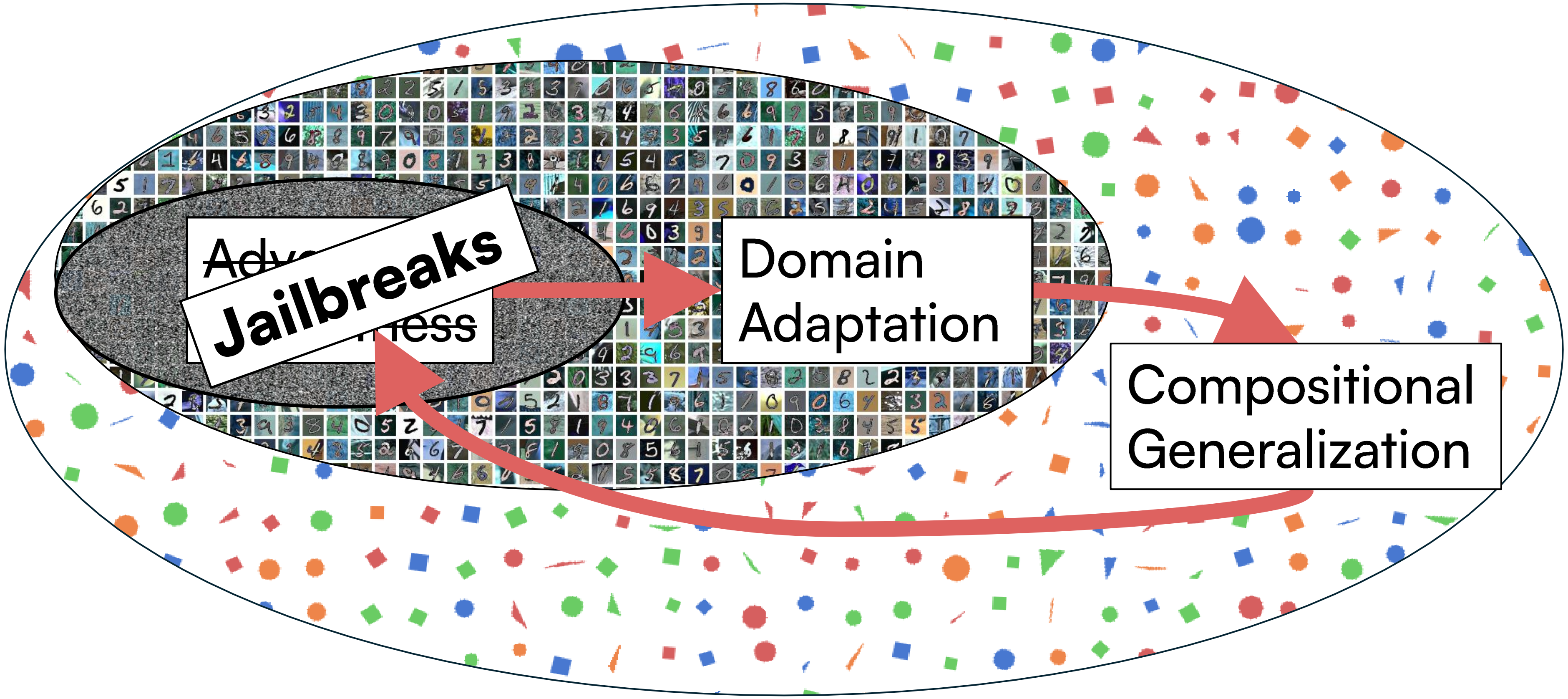
Disclaimer



Disclaimer


Adversarial
Robustness
Domain
Adaptation
Compositional
Generalization

Disclaimer



~~Adversarial Robustness~~
Jailbreaks
Domain Adaptation
Compositional Generalization

Background

On the Binding Problem in Artificial Neural Networks

Klaus Greff* KLAUSG@GOOGLE.COM
Google Research, Brain Team
Tucholskystraße 2, 10116 Berlin, Germany

Sjoerd van Steenkiste SJOERD@IDSIA.CH

Jürgen Schmidhuber JUERGEN@IDSIA.CH
Istituto Dalle Molle di studi sull'intelligenza artificiale (IDSIA)
Università della Svizzera Italiana (USI)
Scuola universitaria professionale della Svizzera italiana (SUPSI)
Via la Santa 1, 6962 Viganello, Switzerland

Abstract

Contemporary neural networks still fall short of human-level generalization, which extends far beyond our direct experiences. In this paper, we argue that the underlying cause for this shortcoming is their inability to dynamically and flexibly bind information that is distributed throughout the network. This *binding problem* affects their capacity to acquire a compositional understanding of the world in terms of symbol-like entities (like objects), which is crucial for generalizing in predictable and systematic ways. To address this issue, we propose a unifying framework that revolves around forming meaningful entities from unstructured sensory inputs (segregation), maintaining this separation of information at a representational level (representation), and using these entities to construct new inferences, predictions, and behaviors (composition). Our analysis draws inspiration from a wealth of research in neuroscience and cognitive psychology, and surveys relevant mechanisms from the machine learning literature, to help identify a combination of inductive biases that allow symbolic information processing to emerge naturally in neural networks. We believe that a compositional approach to AI, in terms of grounded symbol-like representations, is of fundamental importance for realizing human-level generalization, and we hope that this paper may contribute towards that goal as a reference and inspiration.

Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. "On the binding problem in artificial neural networks." arXiv preprint arXiv:2012.05208 (2020).

The Binding Problem

- NNs fall short on OOD, where humans don't

Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. "On the binding problem in artificial neural networks." arXiv preprint arXiv:2012.05208 (2020).

The Binding Problem

- NNs fall short on OOD, where humans don't
- One explanation — NNs learn surface statistics, not underlying concepts

Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. "On the binding problem in artificial neural networks." arXiv preprint arXiv:2012.05208 (2020).

The Binding Problem

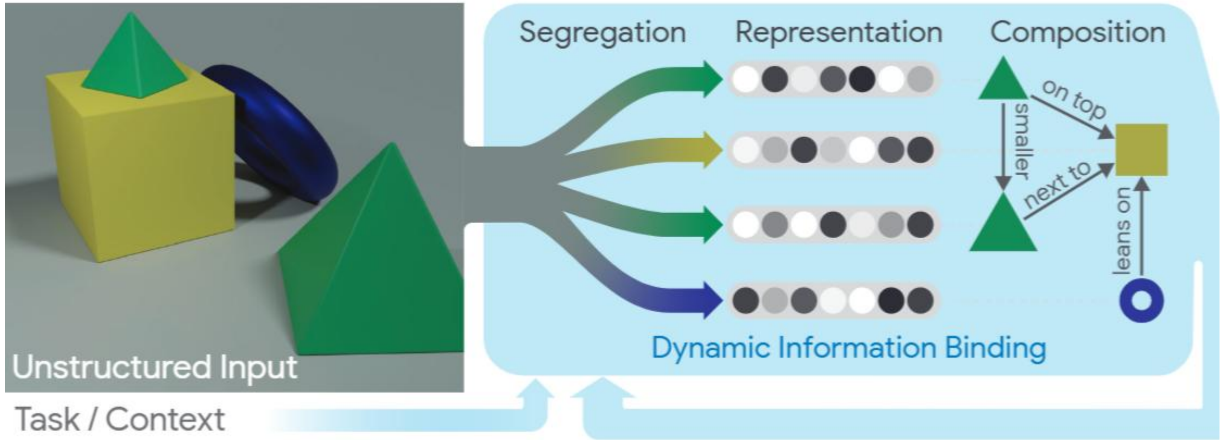
- NNs fall short on OOD, where humans don't
- One explanation — NNs learn surface statistics, not underlying concepts

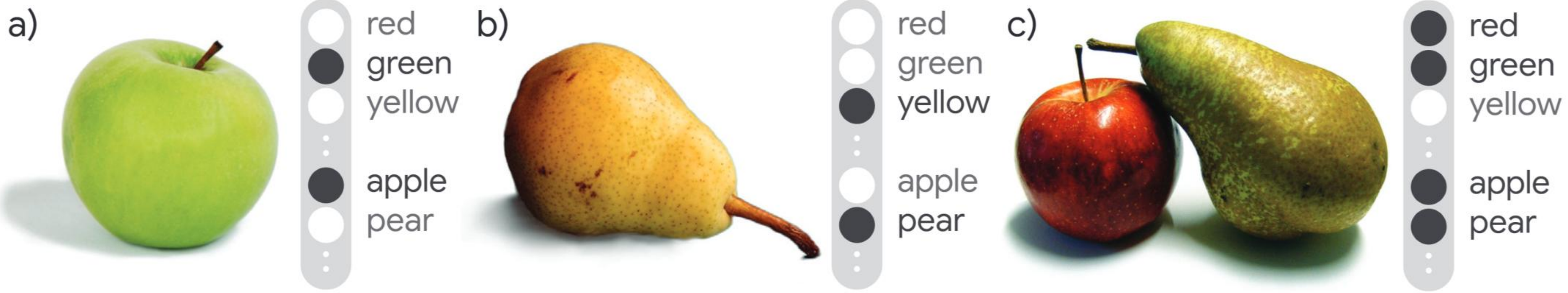
- Toddlers/monkeys/cats exhibit symbolic or object reasoning

Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. "On the binding problem in artificial neural networks." arXiv preprint arXiv:2012.05208 (2020).

The Binding Problem

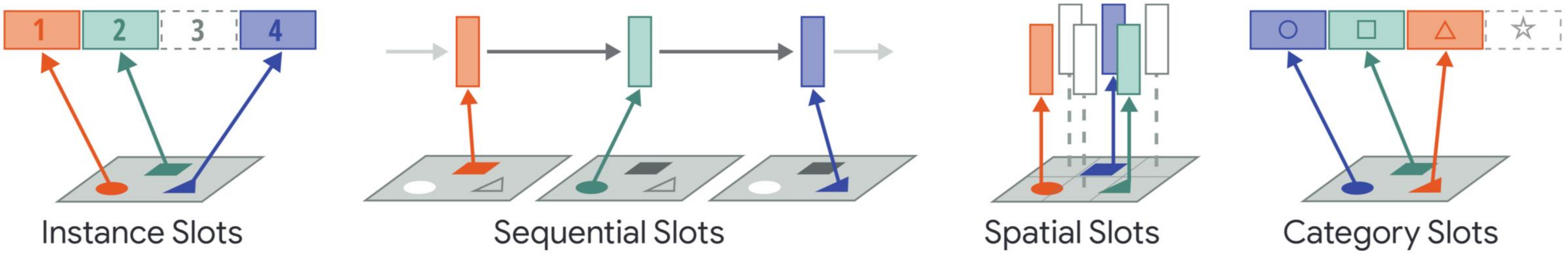
- Object-level abstractions are self-contained
- Idea: add an implicit bias to learn object level abstractions





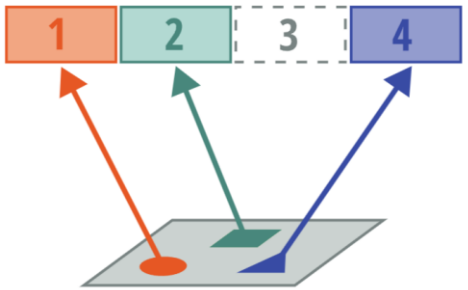
Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. "On the binding problem in artificial neural networks." arXiv preprint arXiv:2012.05208 (2020).

Slots



Greff, Klaus, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. "On the binding problem in artificial neural networks." arXiv preprint arXiv:2012.05208 (2020).

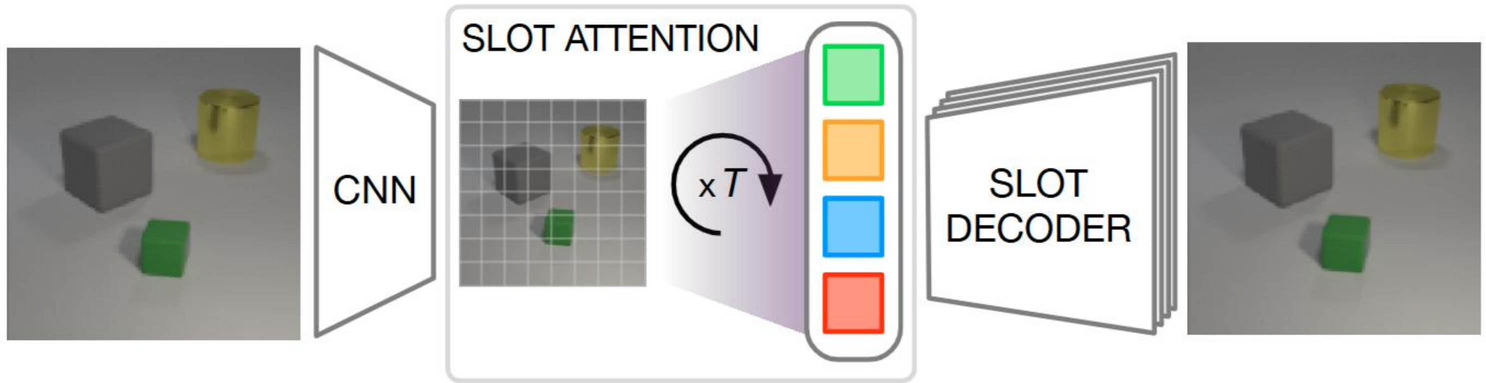
Slots



Instance Slots

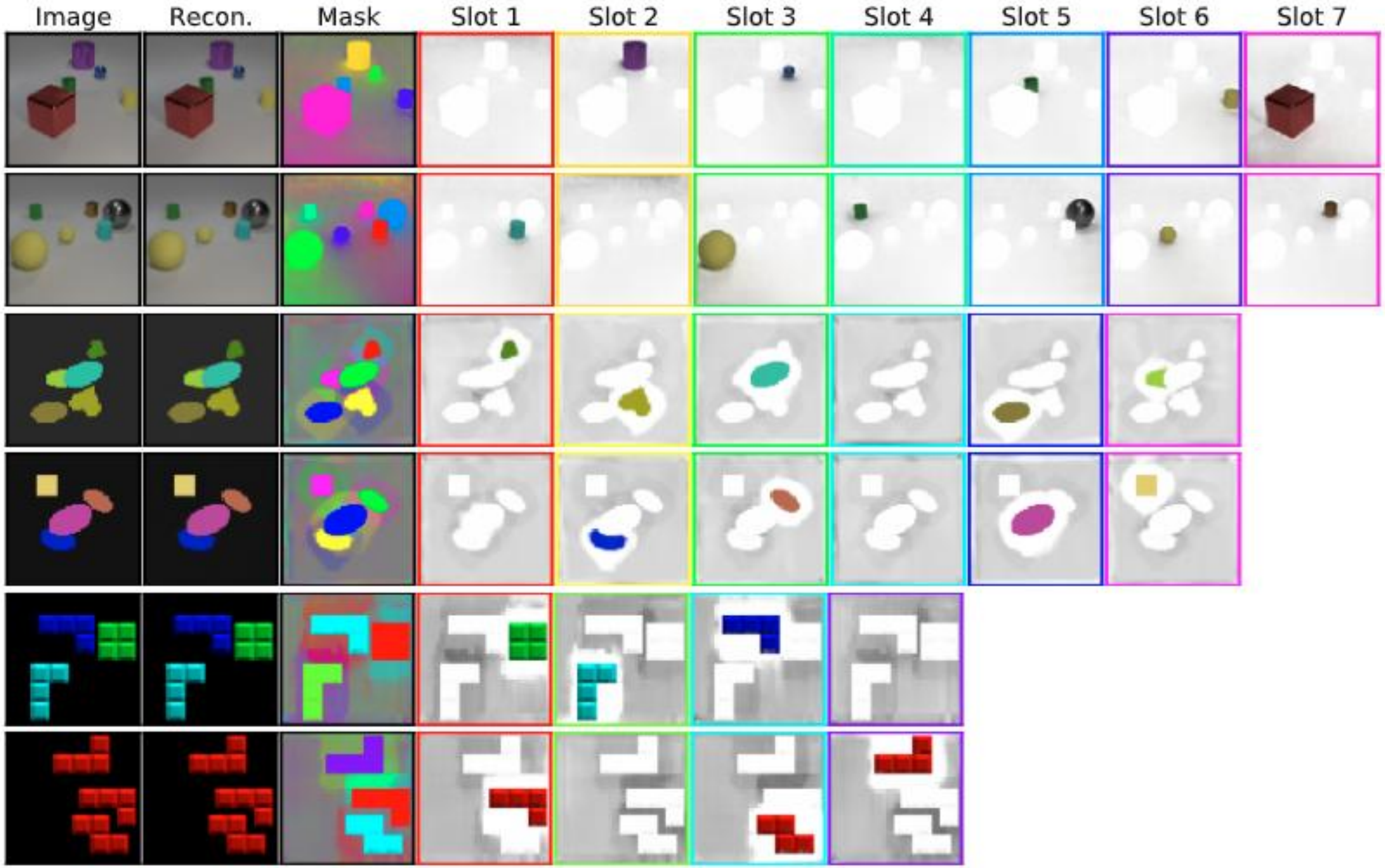
Architectures:

- MONet
  (Burgess, Christopher P., et al, 2019)
- GENESIS
  (Engelcke, Martin, et al, 2019)
- SlotAttention
  (Locatello, Francesco, et al., 2020)
- ...



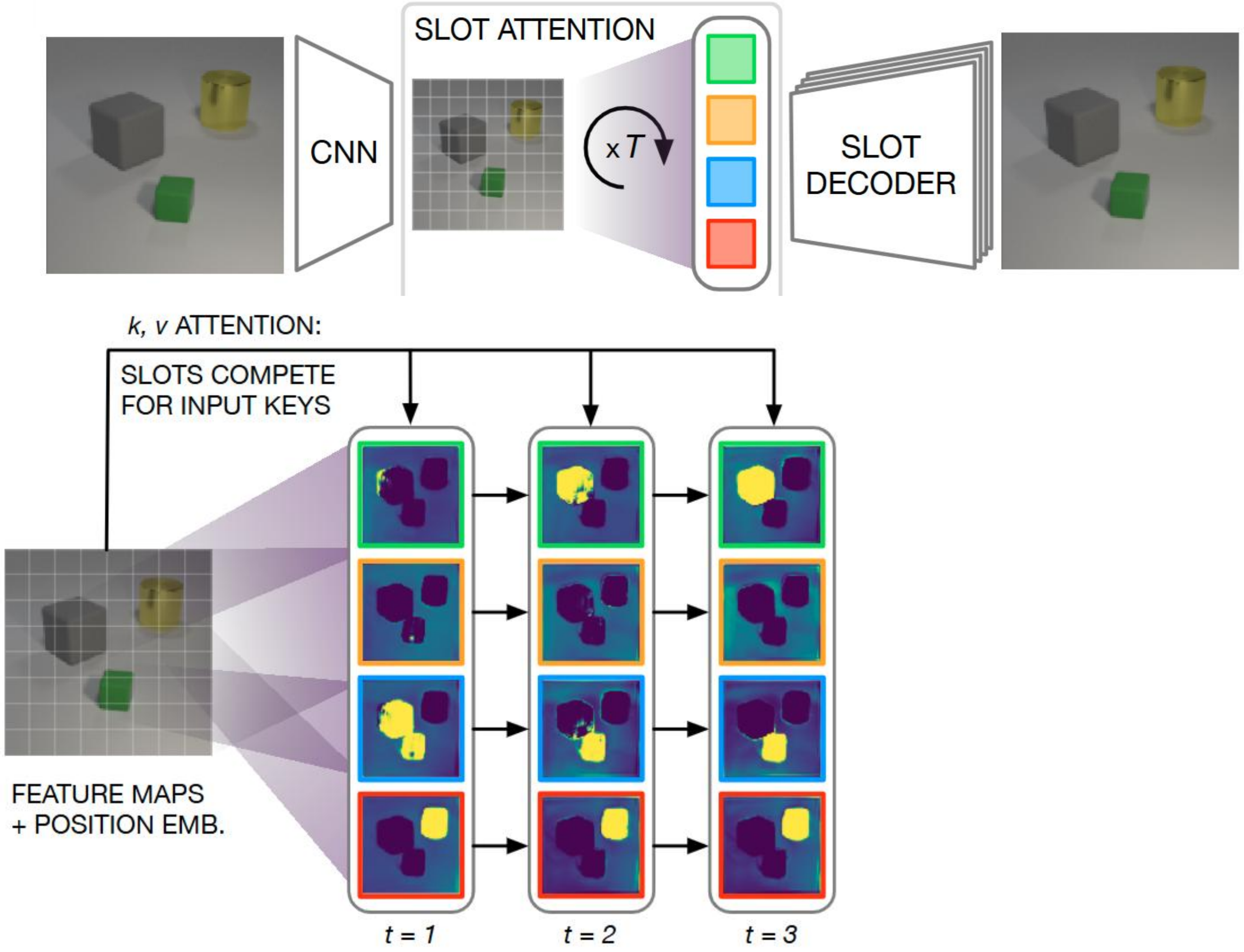
Locatello, Francesco, et al. "Object-centric learning with slot attention." Advances in neural information processing systems 33 (2020): 11525-11538.

Slots



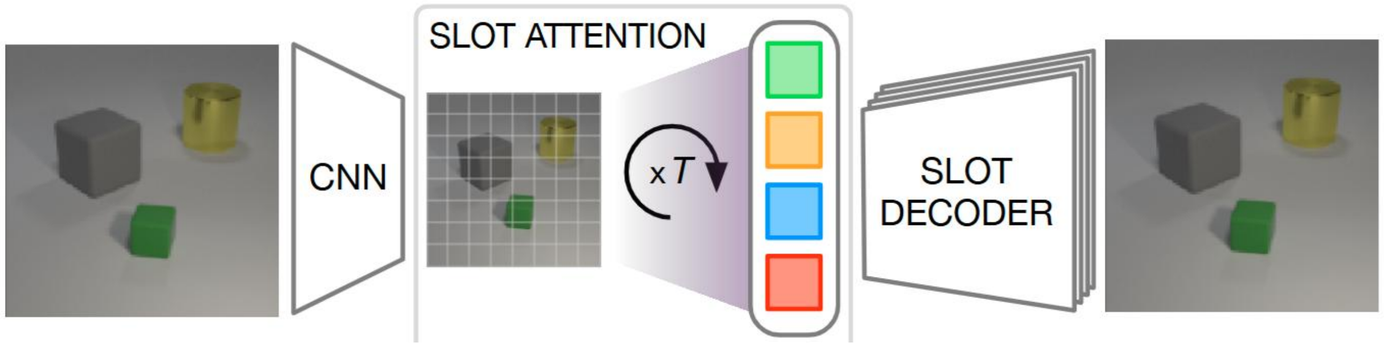
Locatello, Francesco, et al. "Object-centric learning with slot attention." Advances in neural information processing systems 33 (2020): 11525-11538.

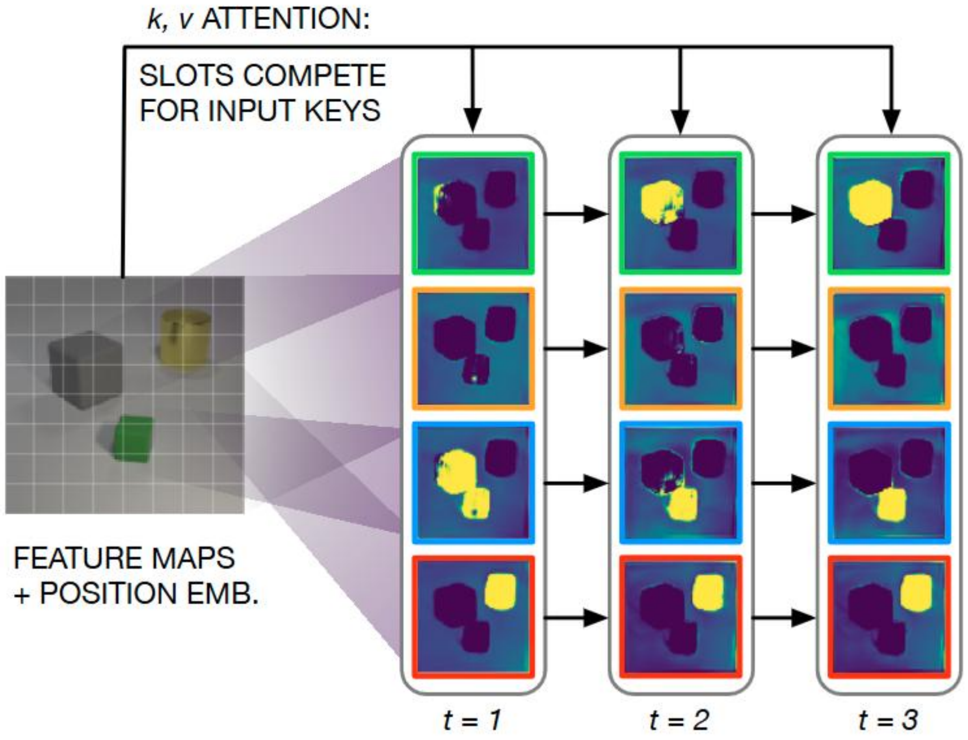
SlotAttention (Locatello, Francesco, et al., 2020)



Locatello, Francesco, et al. "Object-centric learning with slot attention." Advances in neural information processing systems 33 (2020): 11525-11538.

SlotAttention (Locatello, Francesco, et al., 2020)





Algorithm 1 Slot Attention module. The input is a set of N vectors of dimension $D_{\texttt{inputs}}$ which is mapped to a set of K slots of dimension $D_{\texttt{slots}}$. We initialize the slots by sampling their initial values as independent samples from a Gaussian distribution with shared, learnable parameters $\mu \in \mathbb{R}^{D_{\texttt{slots}}}$ and $\sigma \in \mathbb{R}^{D_{\texttt{slots}}}$. In our experiments we set the number of iterations to $T = 3$.

Input: $\texttt{inputs} \in \mathbb{R}^{N \times D_{\texttt{inputs}}}$, $\texttt{slots} \sim \mathcal{N}(\mu,\ \mathrm{diag}(\sigma)) \in \mathbb{R}^{K \times D_{\texttt{slots}}}$
Layer params: $k,\ q,\ v$: linear projections for attention; `GRU`; `MLP`; `LayerNorm` (x3)
  $\texttt{inputs} = \texttt{LayerNorm}(\texttt{inputs})$
  for $t = 0 \dots T$
    $\texttt{slots\_prev} = \texttt{slots}$
    $\texttt{slots} = \texttt{LayerNorm}(\texttt{slots})$
    $\texttt{attn} = \texttt{Softmax}\left(\frac{1}{\sqrt{D}} k(\texttt{inputs}) \cdot q(\texttt{slots})^T,\ \texttt{axis='slots'}\right)$ # norm. over slots
    $\texttt{updates} = \texttt{WeightedMean}(\texttt{weights=attn} + \epsilon,\ \texttt{values=}v(\texttt{inputs}))$ # aggregate
    $\texttt{slots} = \texttt{GRU}(\texttt{state=slots\_prev},\ \texttt{inputs=updates})$ # GRU update (per slot)
    $\texttt{slots} \mathrel{+}= \texttt{MLP}(\texttt{LayerNorm}(\texttt{slots}))$ # optional residual MLP (per slot)
  return `slots`

Locatello, Francesco, et al. "Object-centric learning with slot attention." Advances in neural information processing systems 33 (2020): 11525-11538.

Some examples:
DINOSAUR (Seitzer, Maximilian, et al, 2022)

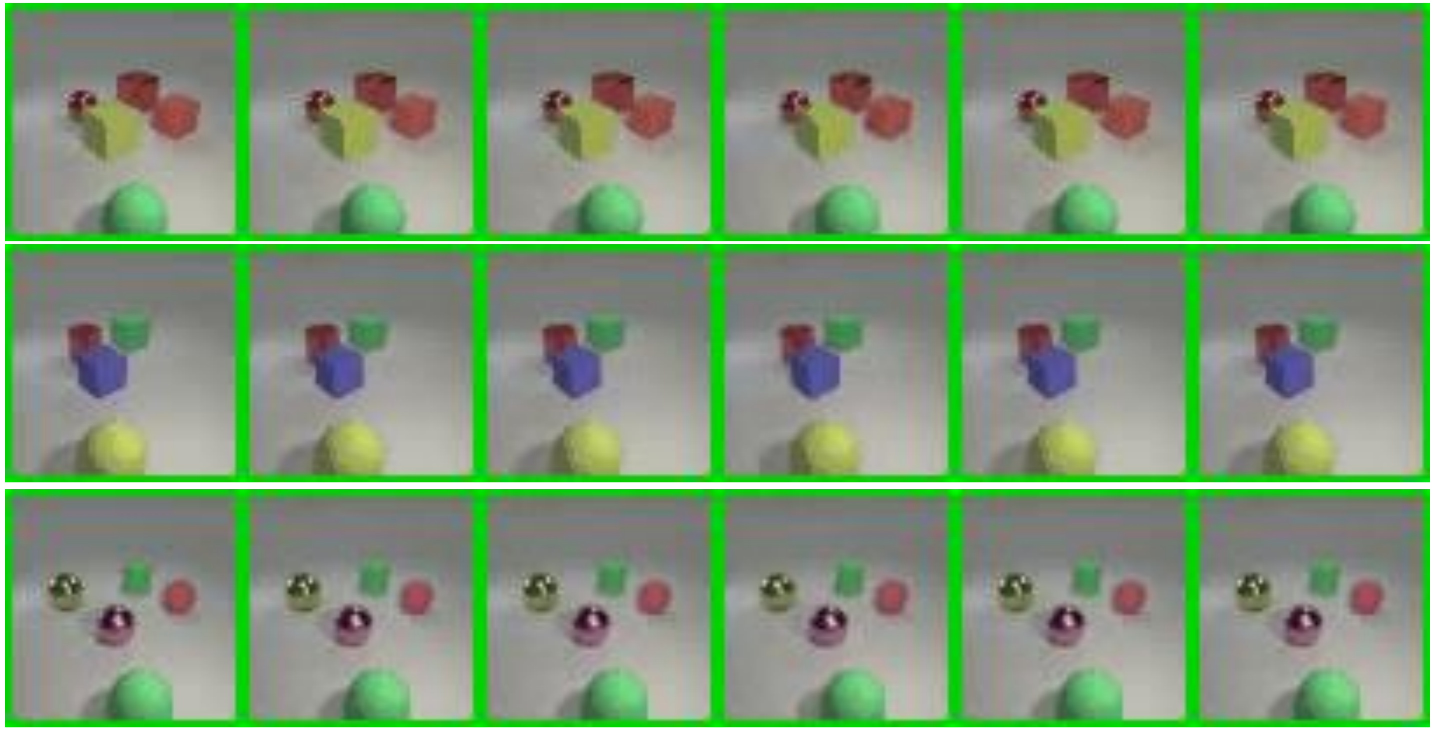
Scaling to “real” data

Seitzer, Maximilian, et al. "Bridging the gap to real-world object-centric learning." arXiv preprint arXiv:2209.14860 (2022).

Some examples:
SlotFormer (Wu, Ziyi, et al, 2022)

Next frame prediction

GT Other methods **SlotFormer**



Wu, Ziyi, et al. "Slotformer: Unsupervised visual dynamics simulation with object-centric models." arXiv preprint arXiv:2210.05861 (2022).

More examples:

Slot-based generation editing, temporal slots, audio slots, etc.

Thomas Kipf

FOLLOW

Other names

Staff Research Scientist, Google DeepMind
Verified email at google.com - Homepage

Machine Learning Graph Representation Lear… Graph Neural Networks

TITLE	CITED BY	YEAR
Neural Assets: 3D-Aware Multi-Object Scene Synthesis with Image Diffusion Models Z Wu, Y Rubanova, R Kabra, DA Hudson, I Gilitschenski, Y Aytar, ... arXiv preprint arXiv:2406.09292	1	2024
Latent Pose Queries for Machine-Learned Image View Synthesis SMM Sajjadi, K Greff, EFR Pot, DC Duckworth, M Lucic, A Mahendran, ... US Patent App. 18/517,190		2024
SceneCraft: An LLM Agent for Synthesizing 3D Scene as Blender Code Z Hu, A Iscen, A Jain, T Kipf, Y Yue, DA Ross, C Schmid, A Fathi International Conference on Machine Learning (ICML)	9	2024
Learning 3D Particle-based Simulators from RGB-D Videos WF Whitney, T Lopez-Guevara, T Pfaff, Y Rubanova, T Kipf, K Stachenfeld, ... International Conference on Learning Representations (ICLR)	4	2024
DyST: Towards Dynamic Neural Scene Representations on Real-World Videos M Seitzer, S van Steenkiste, T Kipf, K Greff, MSM Sajjadi International Conference on Learning Representations (ICLR)	5	2024
DORSal: Diffusion for Object-centric Representations of Scenes $et\ al.$ A Jabri, S van Steenkiste, E Hoogeboom, MSM Sajjadi, T Kipf International Conference on Learning Representations (ICLR)	8	2024

Hopes (Dittadi, Andrea, et al, 2021)

- Object-centric representations are useful for downstream tasks;
- One object's distribution shift does not affect others;
- Object-centric models can still relatively robustly separate objects even under global distribution shifts.

=> Compositional Generalization? Data Efficiency?

Dittadi, Andrea, et al. "Generalization and robustness implications in object-centric learning." arXiv preprint arXiv:2107.00637 (2021).

Provable Compositional Generalization for Object-Centric Learning

Thaddäus Wiedemer[1,2,3*] **Jack Brady**[2,3*] **Alexander Panfilov**[1,2,3*] **Attila Juhos**[1,2,3*]

Matthias Bethge[1,2] **Wieland Brendel**[2,3]

[1]University of Tübingen [2]Tübingen AI Center
[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany

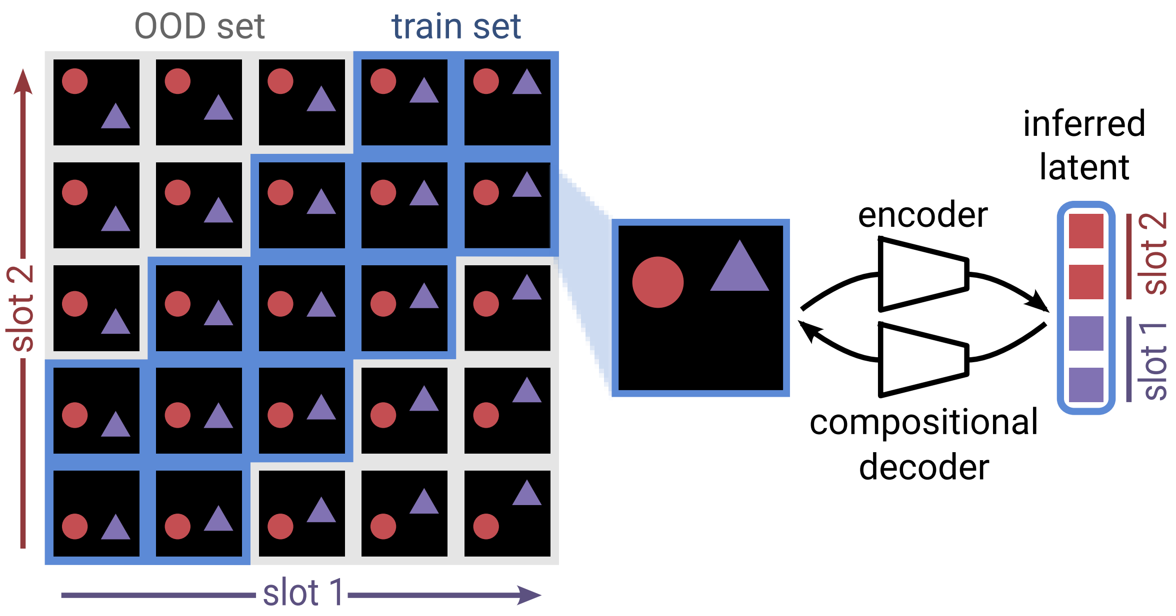
{thaddaeus.wiedemer, jack.brady, attila.juhos}@tue.mpg.de

Abstract

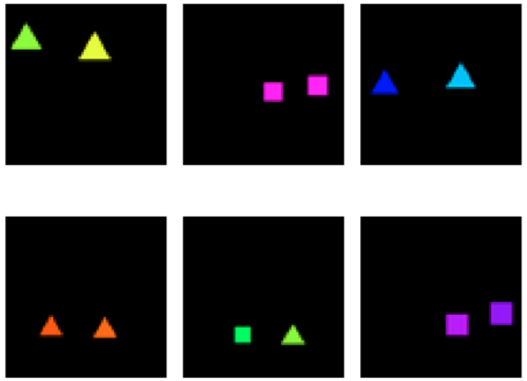
Learning representations that generalize to novel compositions of known concepts is crucial for bridging the gap between human and machine perception. One prominent effort is learning object-centric representations, which are widely conjectured to enable compositional generalization. Yet, it remains unclear when this conjecture will be true, as a principled theoretical or empirical understanding of compositional generalization is lacking. In this work, we investigate when compositional generalization is guaranteed for object-centric representations through the lens of identifiability theory. We show that autoencoders that satisfy structural assumptions on the decoder and enforce encoder-decoder consistency will learn object-centric representations that provably generalize compositionally. We validate our theoretical result and highlight the practical relevance of our assumptions through experiments on synthetic image data.

Wiedemer, Thaddäus, et al. "Provable Compositional Generalization for Object-Centric Learning." arXiv preprint arXiv:2310.05327 (2023).

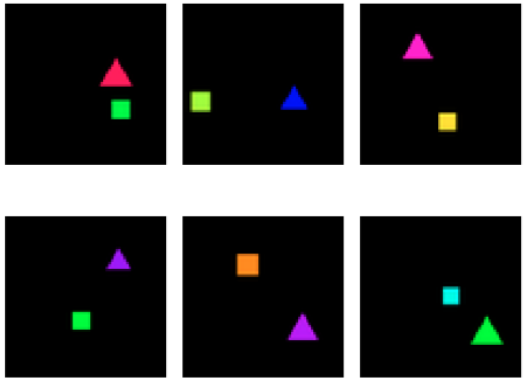
Data & Model



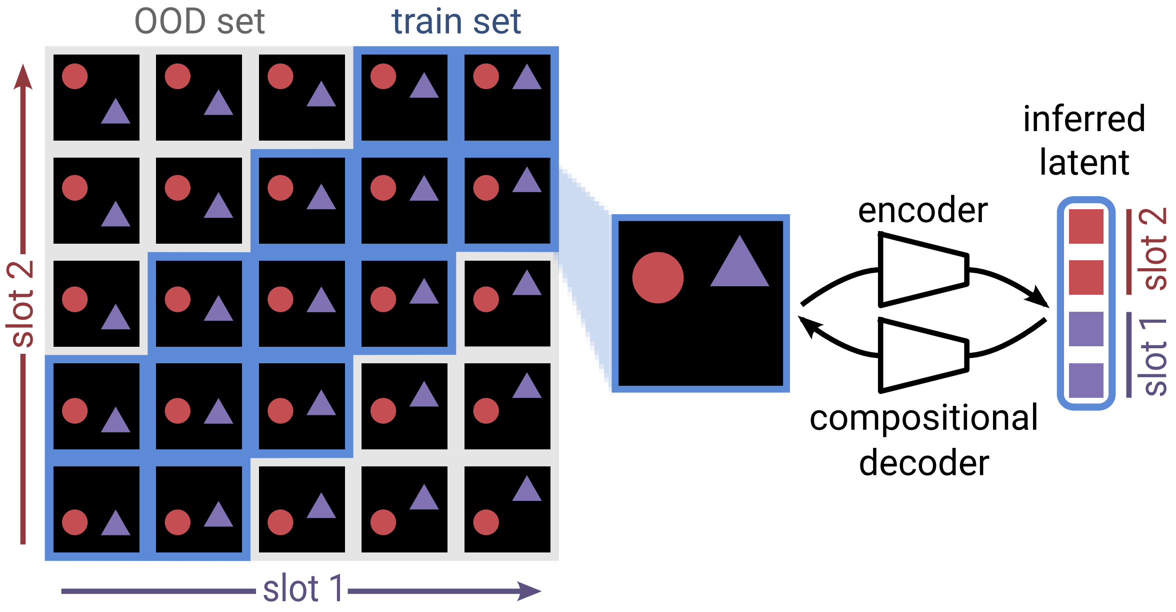
ID set



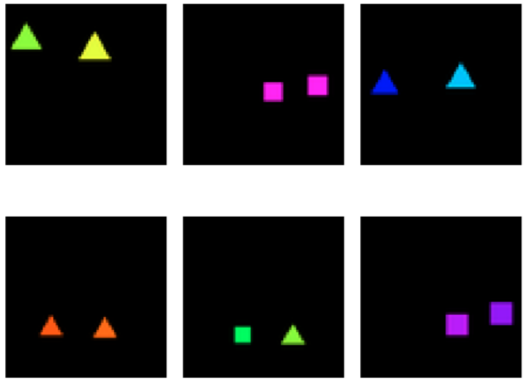
OOD set



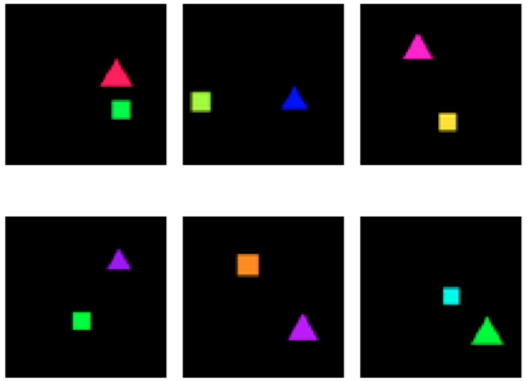
Data & Model



ID set



OOD set



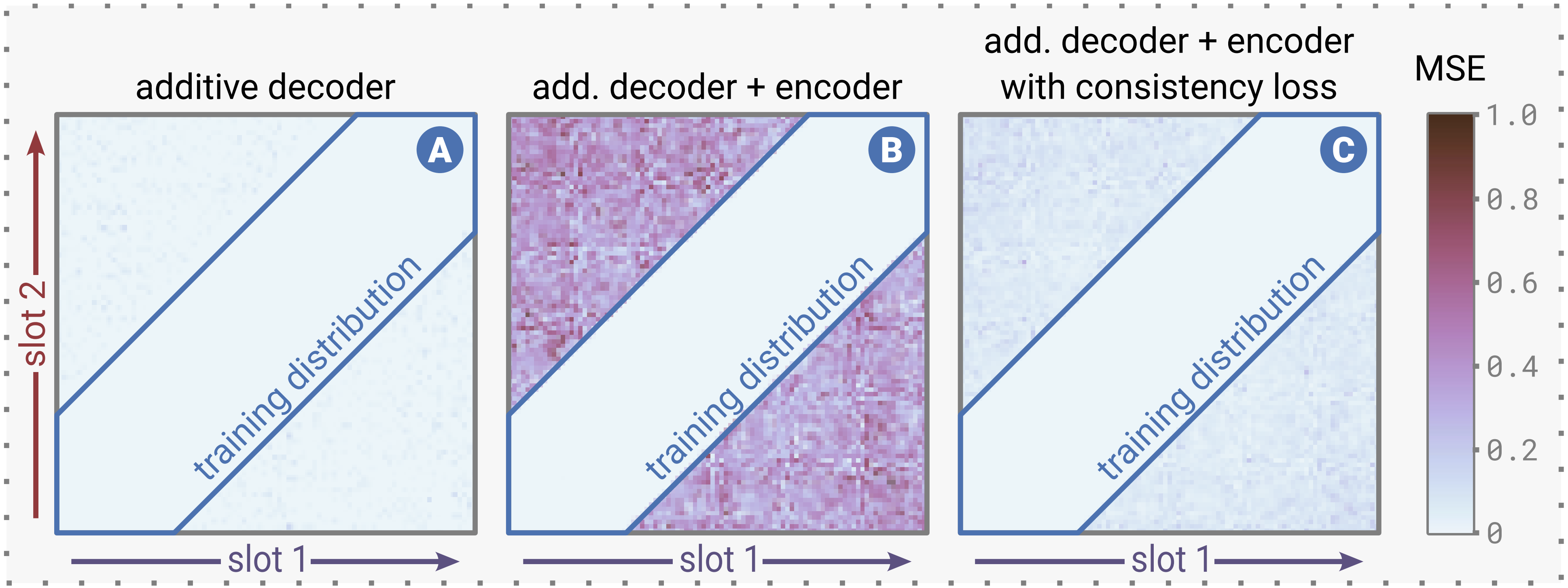
$$g(x) = \hat{z},\ \hat{z} \in \mathbb{R}^{KM}$$

Encoder

$$\hat{x} = \sum_{k=1}^{K} f(\hat{z}_k)$$

Decoder

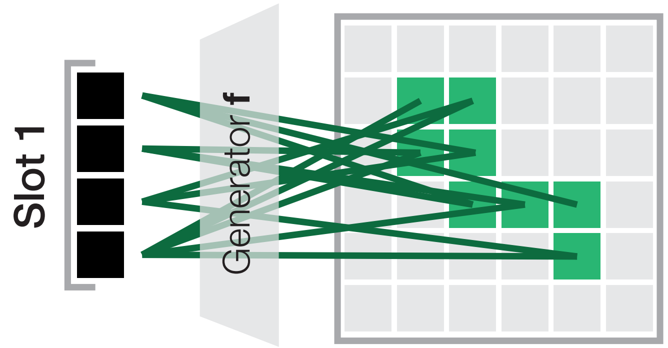
Encoder’s Failure



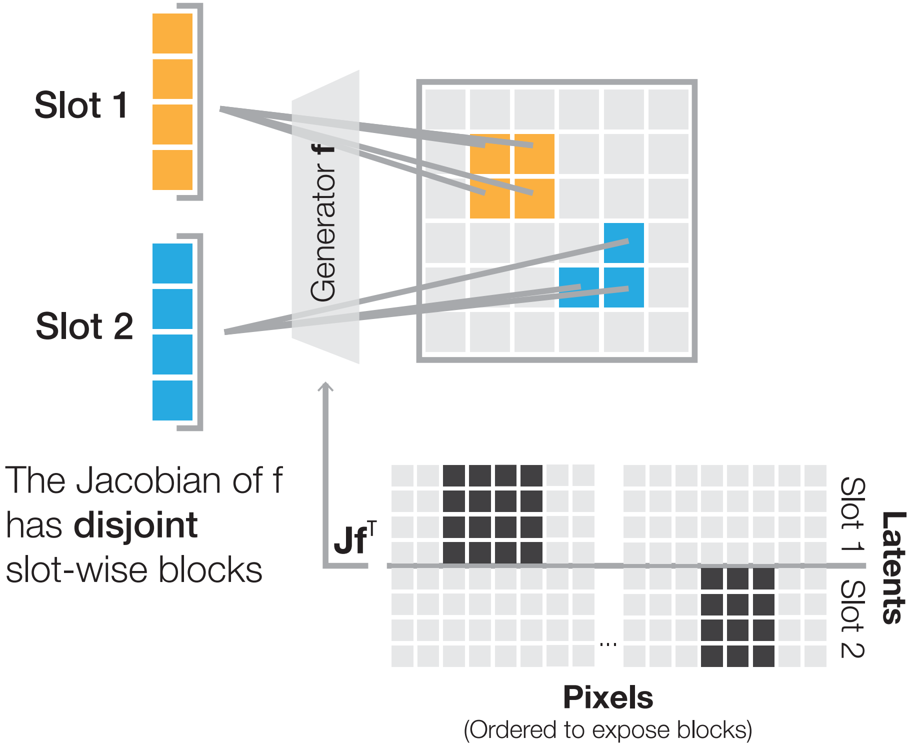
Decoder’s Success

- (Brady, Jack, et al, 2023): **Compositional, irreducible** and **invertable** decoder gives *slot-identifiability*

C | Irreducible Mechanism
Information does not decompose across any pixel subsets S & S′



A | Compositional Generator
Two slots never affect the same pixel



Brady, Jack, et al. "Provably learning object-centric representations." International Conference on Machine Learning. PMLR, 2023.

Decoder's Success

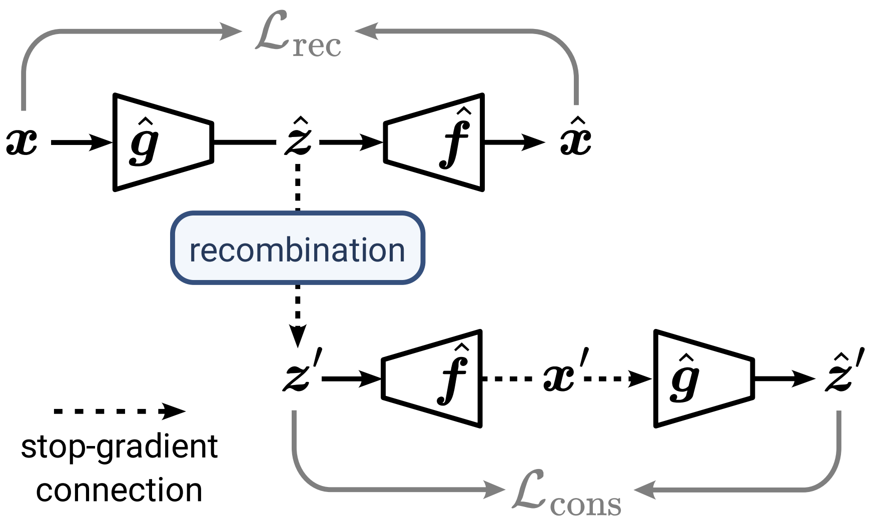
- (Brady, Jack, et al, 2023): **Compositional, irreducible** and **invertable** decoder gives *slot-identifiability*

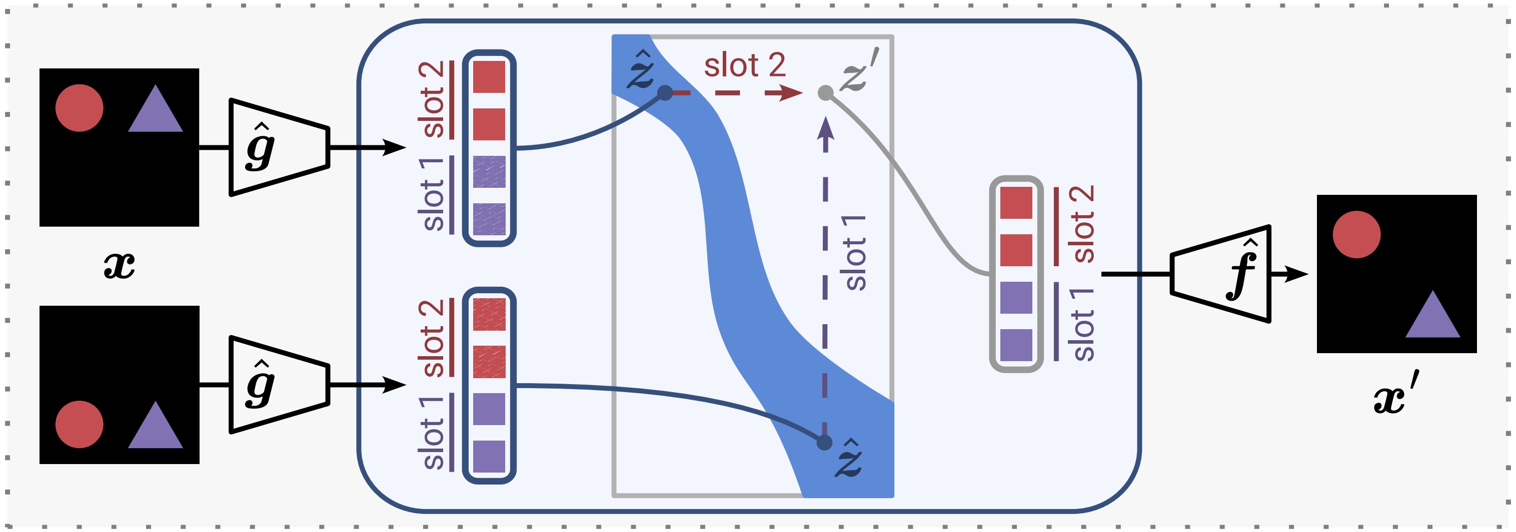
- In Thm. 1 & Thm. 2 we adapted this result, and show that **additive** decoder allows *slot-identifiability* (concurrent to Lachapelle, Sébastien, et al. 2024)

Lachapelle, Sébastien, et al. "Additive decoders for latent variables identification and cartesian-product extrapolation." Advances in Neural Information Processing Systems 36 (2024).

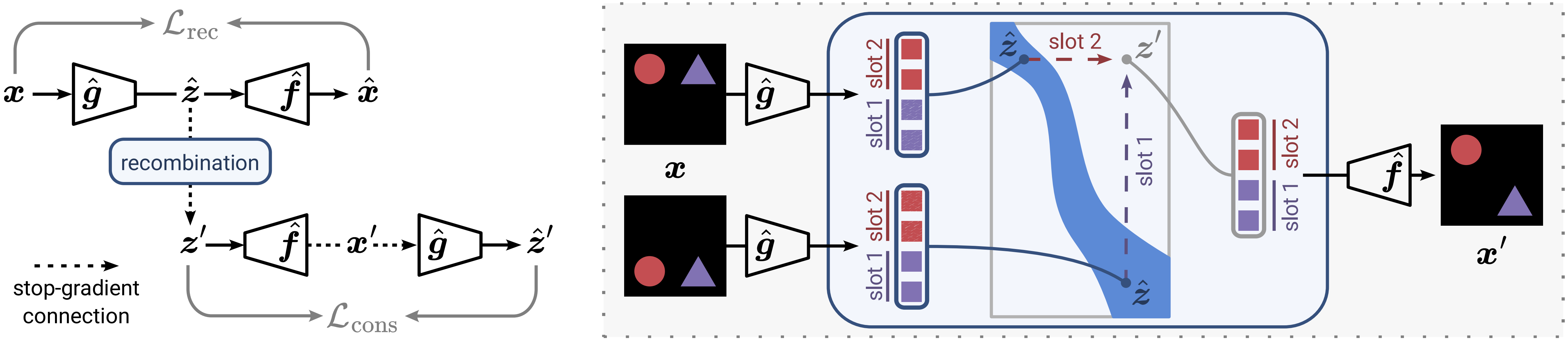
Brady, Jack, et al. "Provably learning object-centric representations." International Conference on Machine Learning. PMLR, 2023.

Aligning Decoder and Encoder



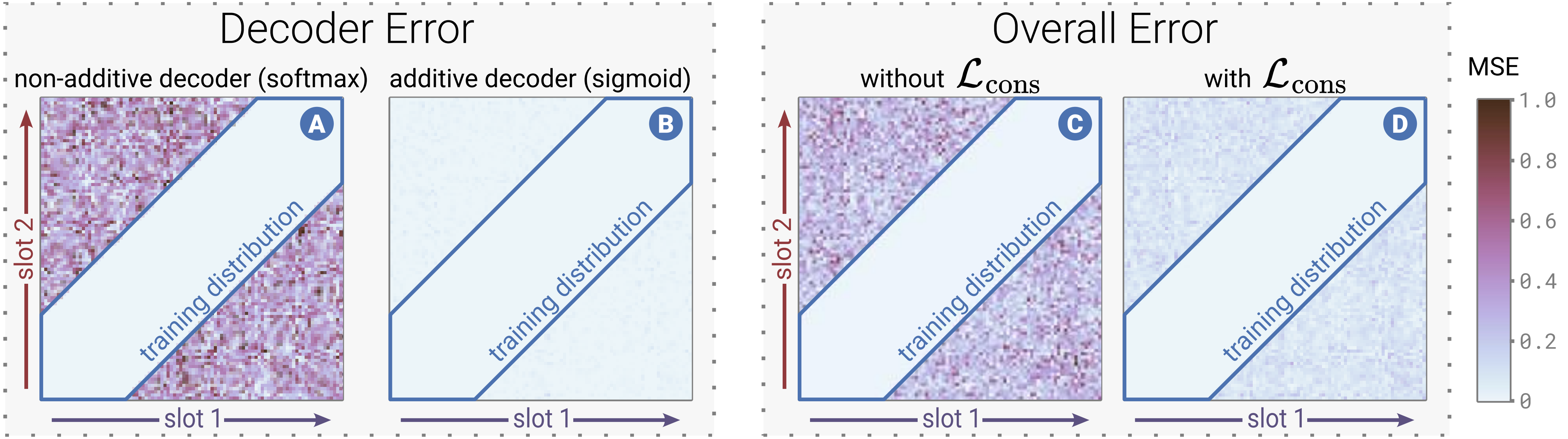


Aligning Decoder and Encoder



In Thm. 3 we show that minimizing both $\mathcal{L}_{\mathrm{rec}}$ and $\mathcal{L}_{\mathrm{cons}}$ on ID combinations allows autoencoder to generalize compositionally

Revisiting SlotAttenion



			Identifiability $R^2\uparrow$		Reconstruction $R^2\uparrow$	
Add.	$\mathcal{L}_{\text{cons}}$	**Det.**	**ID**	**OOD**	**ID**	**OOD**
✗	✗	✗	0.99 $_{\pm 1.7e-3}$	0.81 $_{\pm 9.0e-2}$	0.99 $_{\pm 1.0e-4}$	0.71 $_{\pm 1.9e-2}$
✓	✗	✗	0.99 $_{\pm 2.3e-3}$	0.83 $_{\pm 5.4e-2}$	0.99 $_{\pm 5.8e-4}$	0.72 $_{\pm 2.1e-2}$
✓	✓	✗	0.99 $_{\pm 2.9e-2}$	0.92 $_{\pm 3.4e-2}$	0.99 $_{\pm 8.3e-4}$	0.79 $_{\pm 7.2e-2}$
✓	✓	✓	0.99 $_{\pm 1.9e-3}$	**0.94** $_{\pm 2.2e-2}$	0.99 $_{\pm 1.9e-4}$	**0.92** $_{\pm 2.1e-2}$

Takeaways

- In object-centric learning compositional generalization is possible theoretically and empirically;
- Current assumptions for provable generalization are too restrictive for real world data;
- Training on synthetic data can lead to a better generalization;
- It is hard to scale consistency loss to more slots. Input/output normalization is crucial.