Invariant Slot Attention

Northeastern University
Khoury College of Computer Sciences

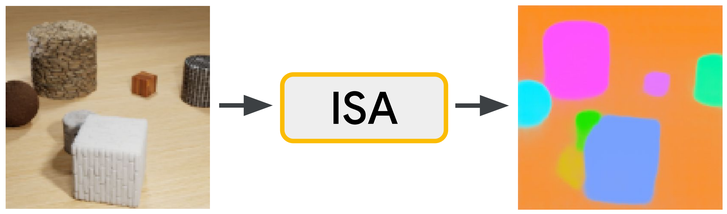
Ondrej Biza[1,*], Sjoerd van Steenkiste[2], Mehdi S. M. Sajjadi[2], Gamaleldin F. Elsayed[2,†], Aravindh Mahendran[2,†], Thomas Kipf[2,†]

[1] Northeastern University. [2] Google Research. [†] Equal contribution. [*] Work performed while at Google. Contact: biza.o@northeastern.edu.

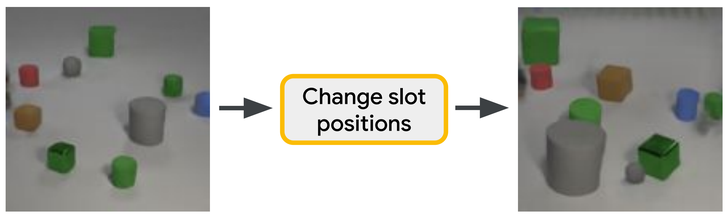
Google Research

Introduction

Can we discover meaningful representations of objects in images without supervision?

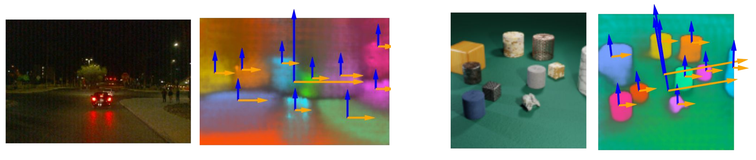


Idea: learn explicit slot poses and scales for higher sample-efficiency and controllability.

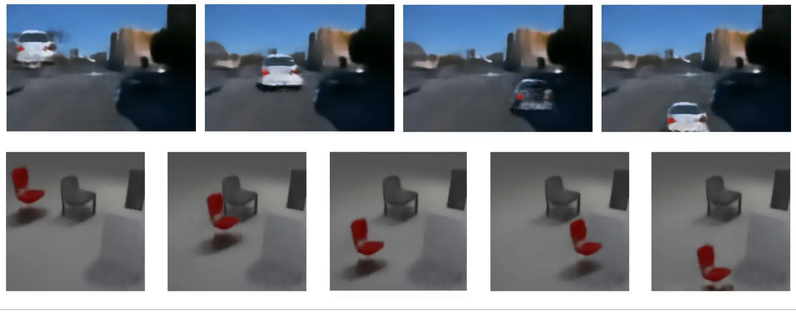


Visualizations

Discovered soft segmentation masks and slot reference frames:

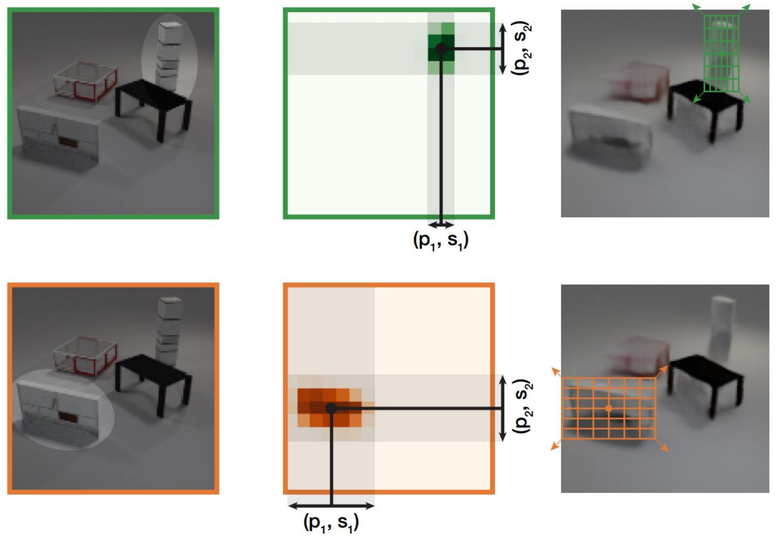


Controlling slot positions:



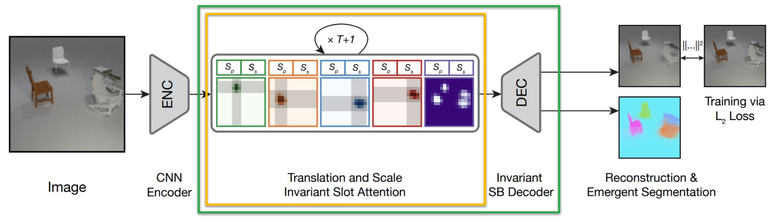
Invariant Slot Attention (ISA)

ISA encodes and decodes objects relative to their poses and scales.



1. Slot attends to an encoded image.
2. Compute slot position and scale from attention masks.
3. Create relative coordinate grids.

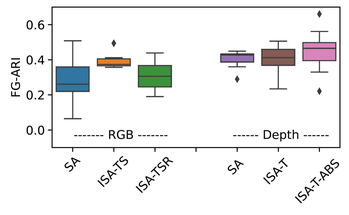
We compute positions, scales and optionally orientations from attention masks. We then use these statistics to create relative coordinate grids in each subsequent round of Slot Attention and in the Spatial Broadcast Decoder. See the full model below:



Quantitative Results

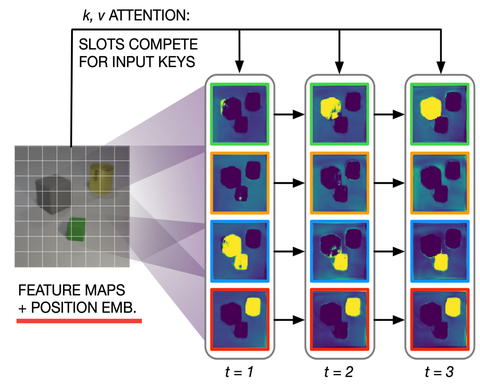
We measure segmentation accuracy (FG-ARI) on CLEVRTex (left) and Waymo Open (right). SA: Slot Attention, ISA: Invariant Slot Attention (ours).

Method	Main	CAMO	OOD
SPACE	17.5 ±4.1	10.6 ±2.1	12.7 ±3.4
DTI	79.9 ±1.4	72.9 ±1.9	73.7 ±1.0
AST-Seg-B3-CT	94.8 ±0.5	87.3 ±3.8	83.1 ±0.8
SA (CNN)	54.5 ±1.6	53.0 ±1.6	54.2 ±2.6
ISA-T (CNN)	66.8 ±5.7	65.0 ±4.9	65.1 ±4.8
ISA-TS (CNN)	78.8 ±3.9	72.9 ±3.5	73.2 ±3.1
SA (ResNet)	91.3 ±2.7	84.9 ±2.9	81.4 ±1.4
ISA-T (ResNet)	87.4 ±6.6	79.0 ±5.9	78.6 ±4.9
ISA-TS (ResNet)	92.9 ±0.4	86.2 ±0.8	84.4 ±0.8



Background: Slot Attention

- Slot-based model – encodes images into set of slots.
- Slots compete over pixels.
- Equivariant to permutation of slots.
- Sensitive to absolute positions of pixels.



Object-Centric Learning with Slot Attention. Locatello et al. NeurIPS 2020.

Conclusion

ISA: Object Discovery with Slot-Centric Reference Frames

- Slots have 2D poses and scales.
- Higher sample-efficiency.
- Controllable slots.

Limitations:

- Exact spatial symmetries are broken by lighting, occlusions, etc.
- Difficult to fit complex real-world data from scratch.

Future work:

- Slots with 3D poses and scales.
- Using pre-trained backbones.
- Invariance in Detection Transformers.

Demo

Website (paper, code, weights, demo)