Spatial Symmetry in Slot Attention

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.

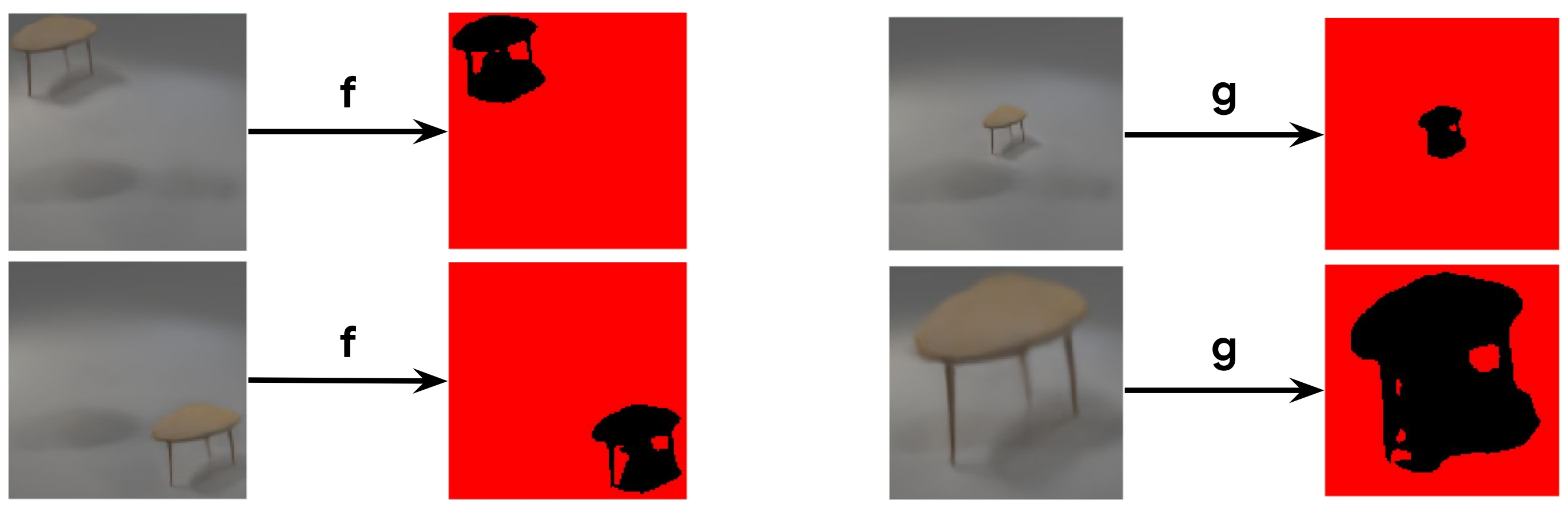
[*] Work performed while at Google. Contact: biza.o@northeastern.edu.

Google Research

Northeastern University
Khoury College of
Computer Sciences

Introduction

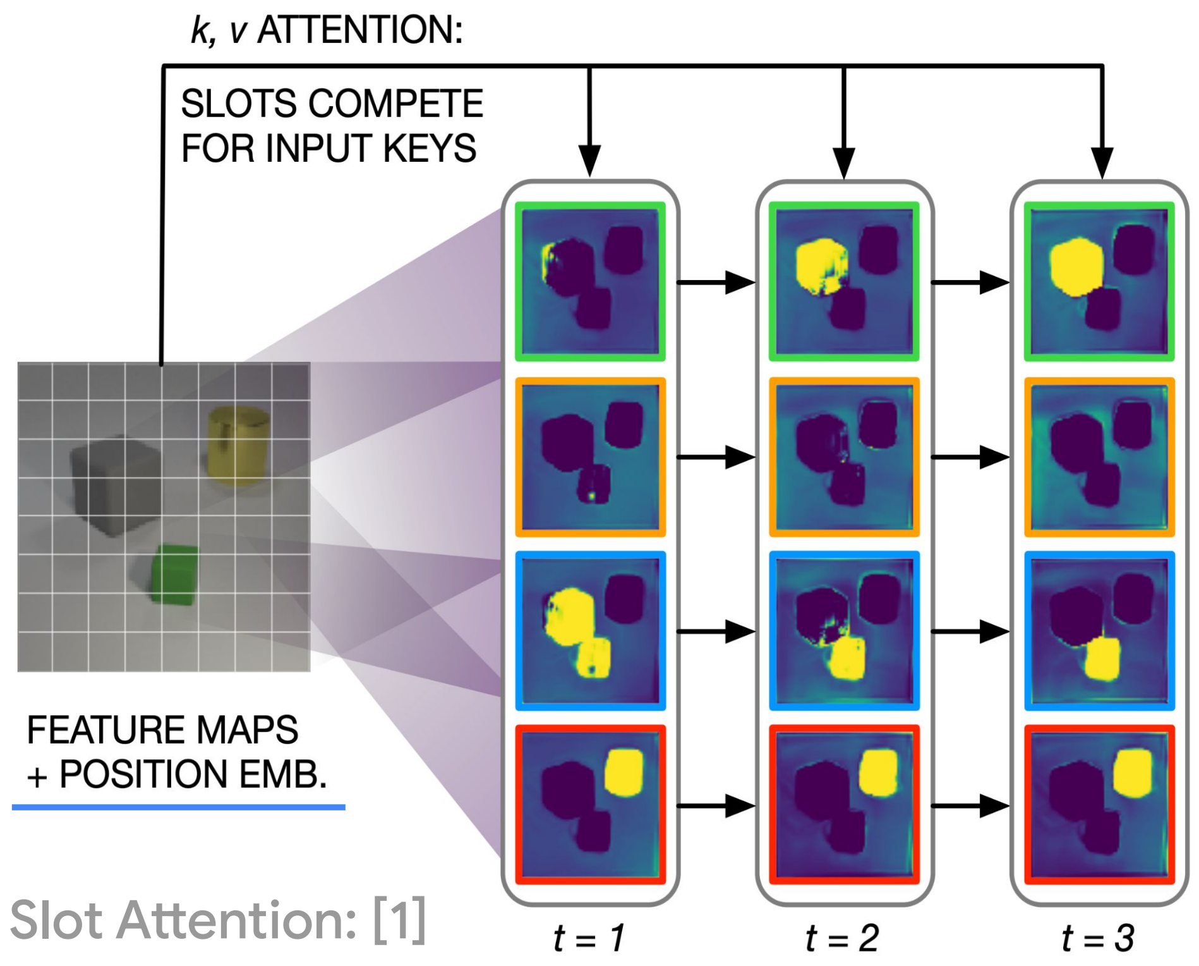
Our goal is to discover objects in scenes without supervision. We show that equivariance to translation and scaling is a useful inductive bias.



f is equivariant to translation.

g is equivariant to scaling.

Background: Slot Attention

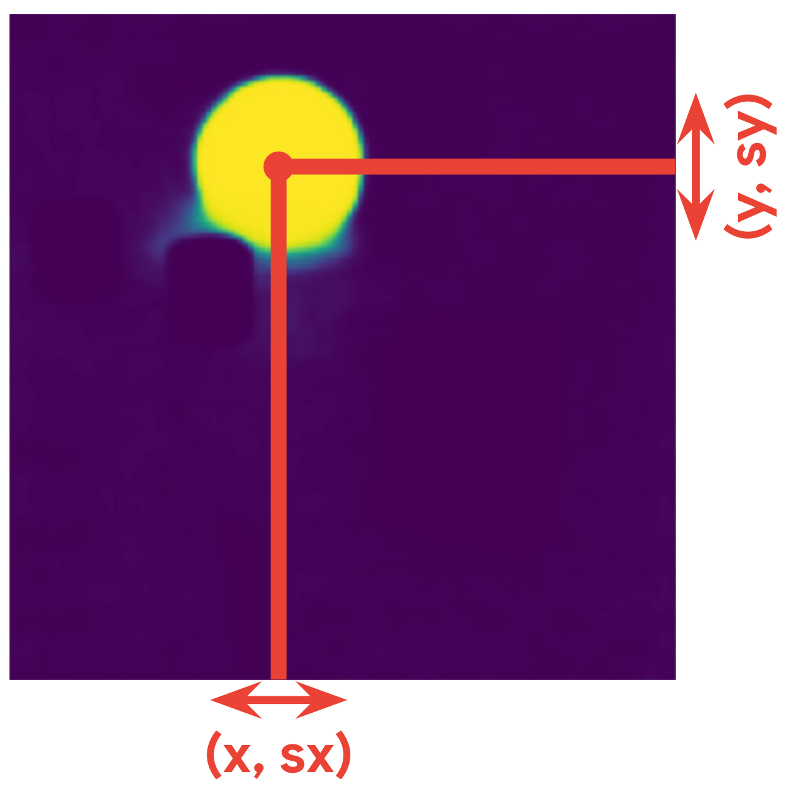


Slot Attention: [1]

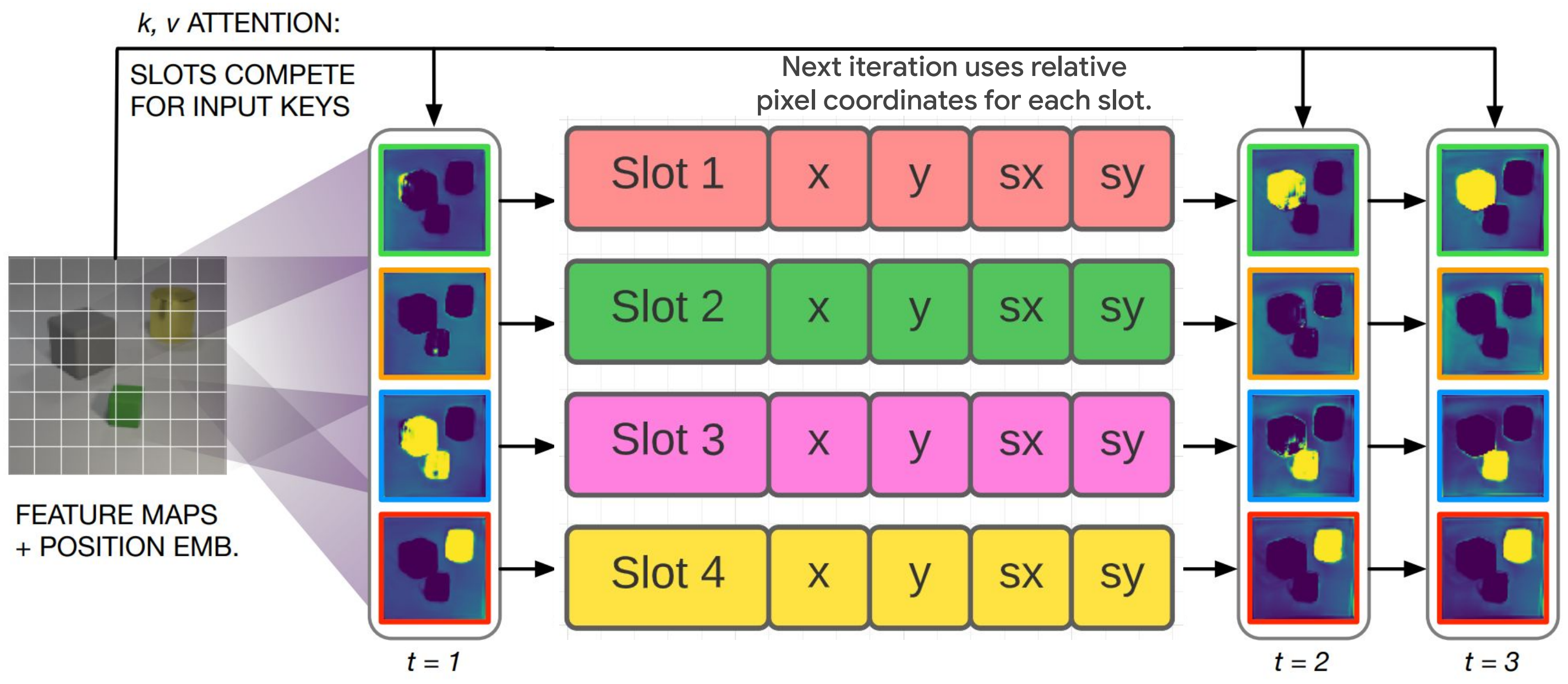
- Learnable clustering for object discovery.
- Equivariant to permutations of objects.
- Sensitive to absolute positions of objects / pixels.

Translation and Scale Equivariant Slot Attention

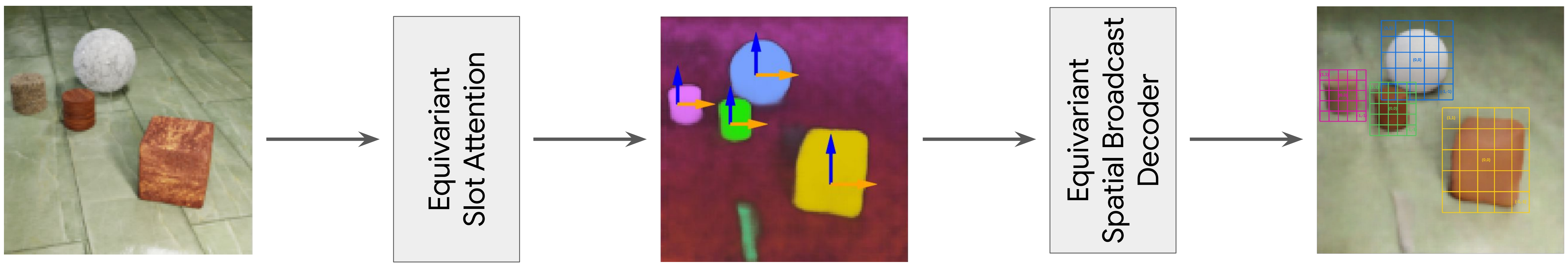
1. Compute slot's position and scale using its attention mask.



2. Equip slots with positions and scales, re-compute at each iteration of Slot Attention.


k, v ATTENTION:
SLOTS COMPETE FOR INPUT KEYS
Next iteration uses relative pixel coordinates for each slot.
FEATURE MAPS + POSITION EMB.
Slot 1 x y sx sy
Slot 2 x y sx sy
Slot 3 x y sx sy
Slot 4 x y sx sy
t = 1
t = 2
t = 3

3. Encode and decode images such that pixel coordinates are represented relative to per-slot reference frames.


Equivariant Slot Attention
Equivariant Spatial Broadcast Decoder

Input image

Predicted segmentation mask with slot-centric reference frames

Decoded image using relative coordinate grids

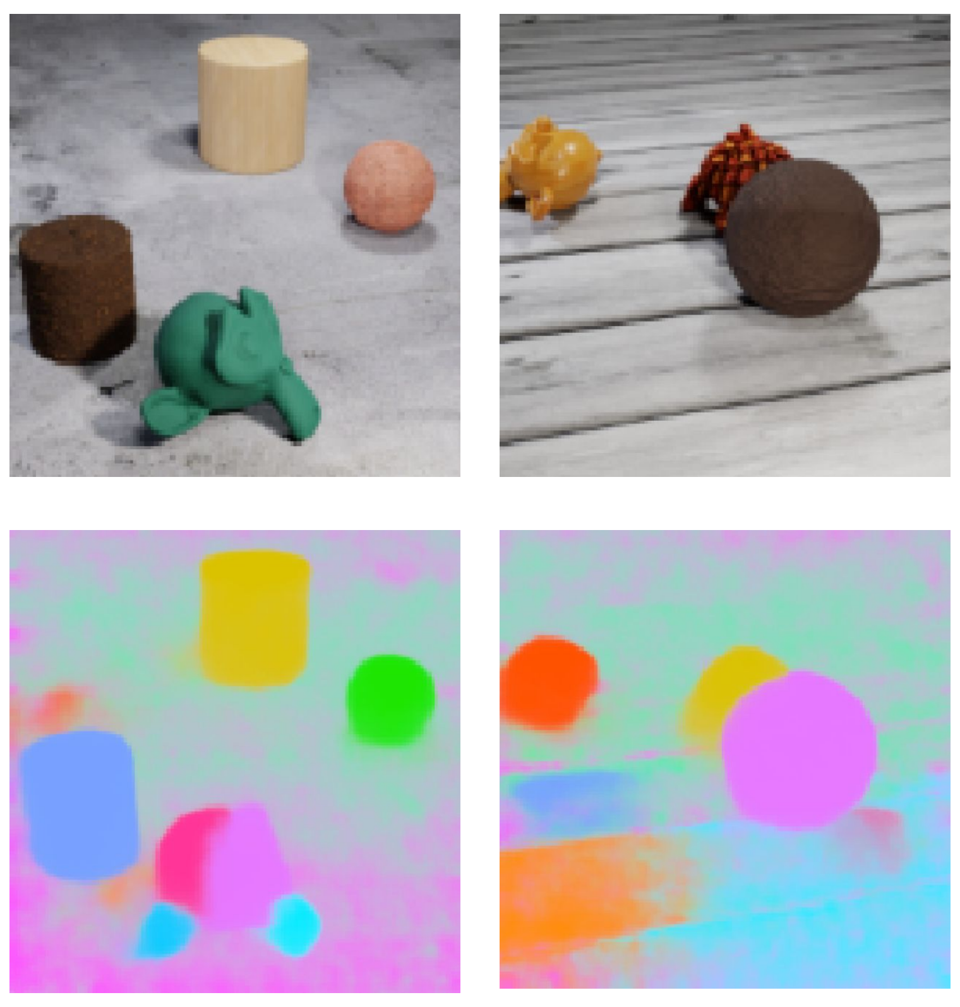
Spatial Broadcast Decoder: [2]

Results

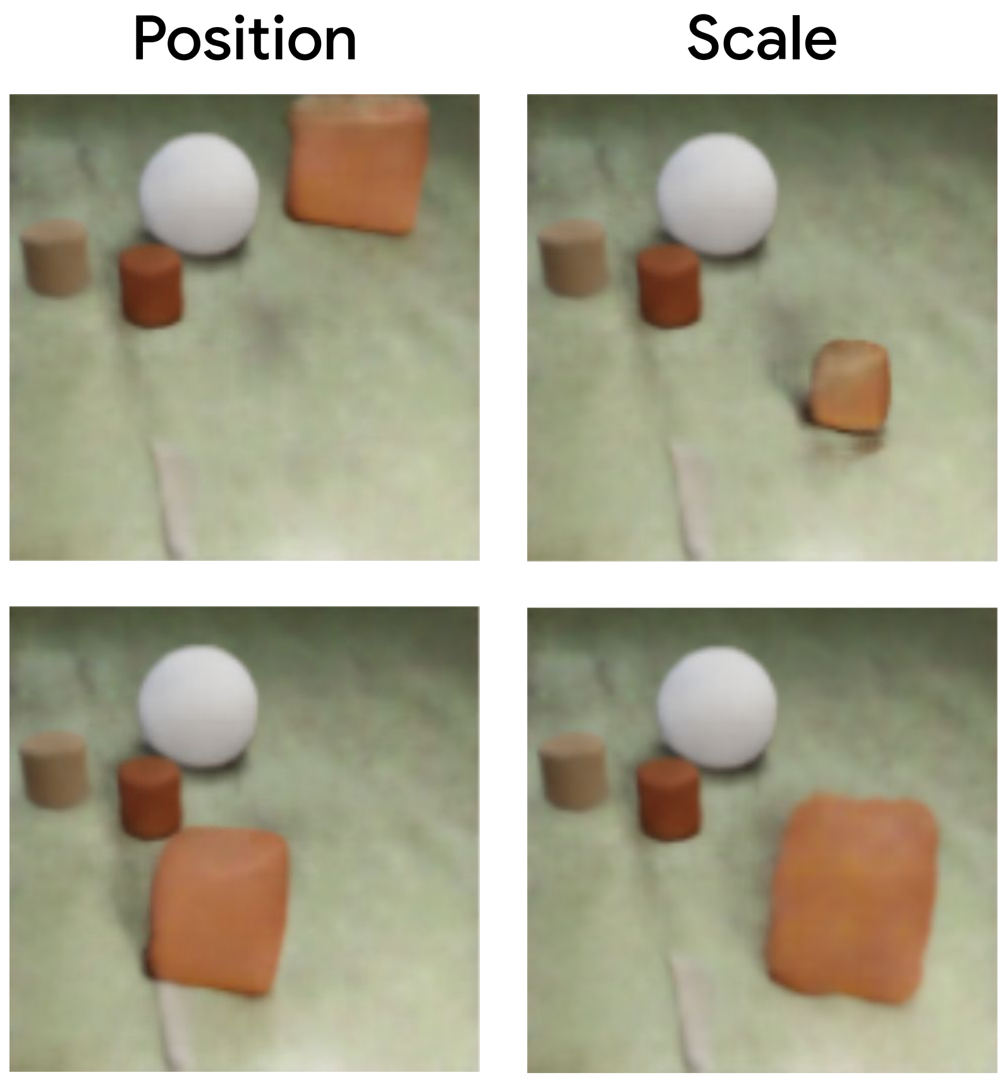
Translation and scaling equivariance in Slot Attention leads to large improvements on a challenging synthetic dataset.

Method	CLEVRTex	
	↑FG-ARI	↓MSE
SimpleCNN SA	$54.5_{\pm 1.6}$	$241_{\pm 14}$
SimpleCNN T-SA	$66.8_{\pm 5.7}$	$230_{\pm 20}$
SimpleCNN TS-SA	$74.1_{\pm 6.4}$	$224_{\pm 4}$
ResNet SA	$80.8_{\pm 12.3}$	$230_{\pm 45}$
ResNet T-SA	$87.6_{\pm 4.0}$	$198_{\pm 21}$
ResNet TS-SA	$86.4_{\pm 9.4}$	$219_{\pm 63}$

Soft segmentation masks.



Editing slots.



[1]: Object-Centric Learning with Slot Attention. Locatello et al. NeurIPS 2020.

[2]: Spatial Broadcast Decoder. Watters et al. arXiv. 2019.