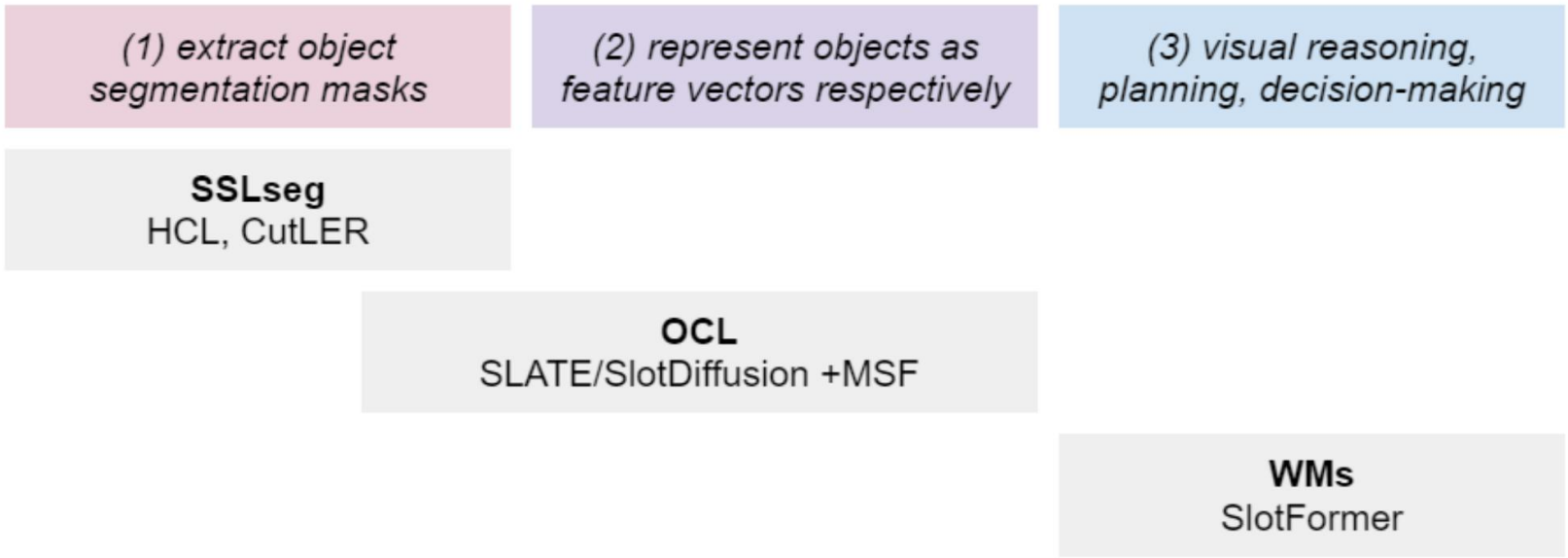
ICLR 2025 Poster

Multi-Scale Fusion for Object Representation

Rongzhen Zhao, Vivienne Wang, Juho Kannala, Joni Pajarinen

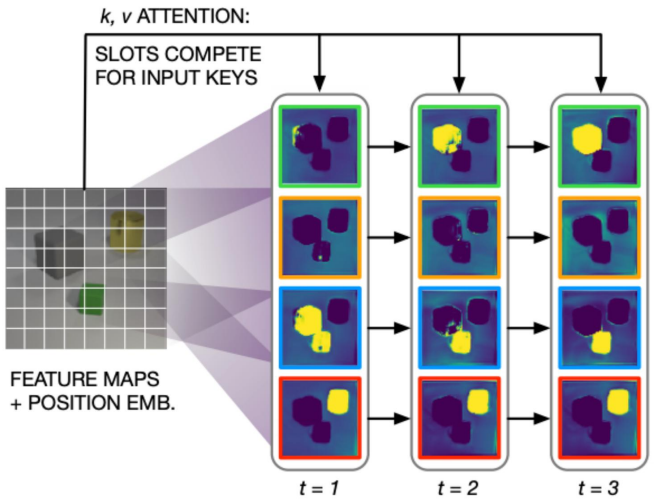
Background

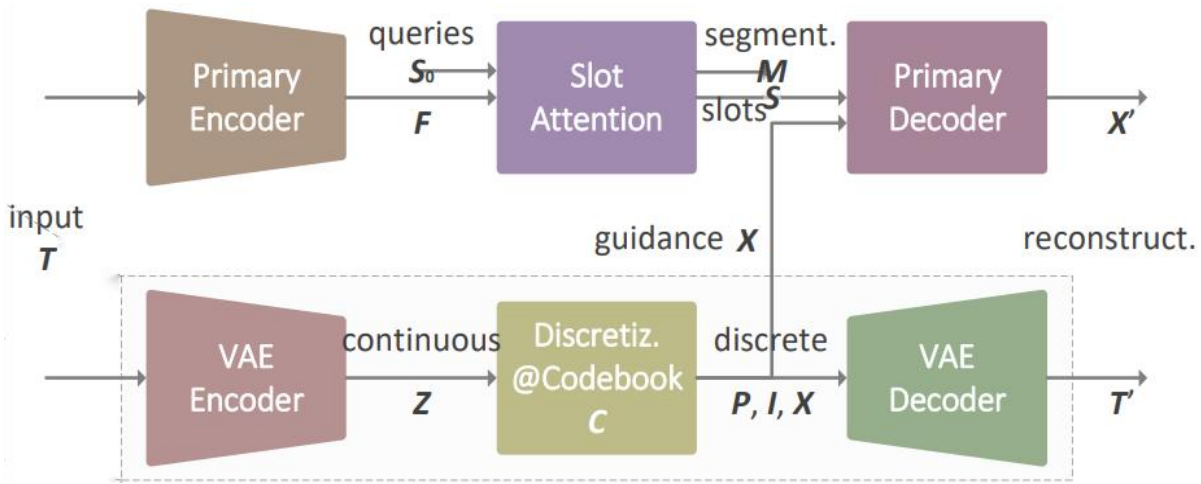
- Representing images or videos as object-level feature vectors, rather than pixel level feature maps, facilitates advanced vision tasks.
  - prediction, reasoning, planning, decision-making



Background

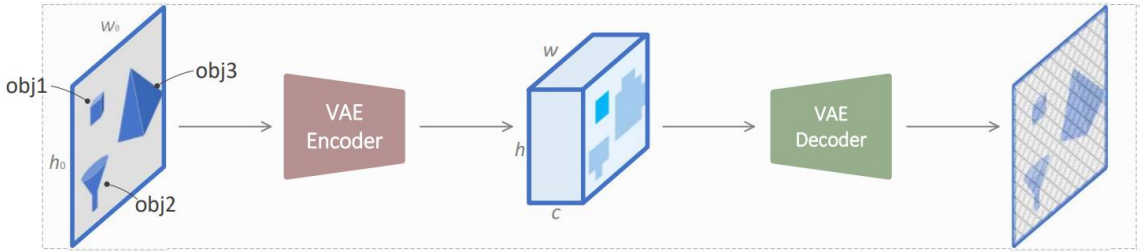
- Object-Centric Learning (OCL) primarily achieves this by
  - reconstructing the input under the guidance of Variational Autoencoder (VAE) representation,
  - to drive *slots* to aggregate as much object information as possible.

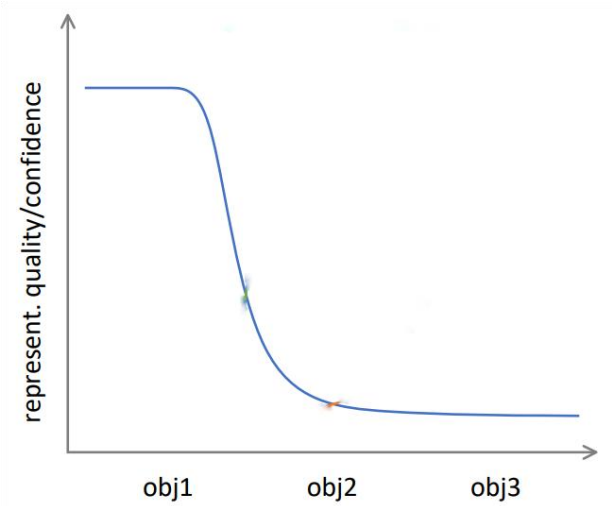




Problem

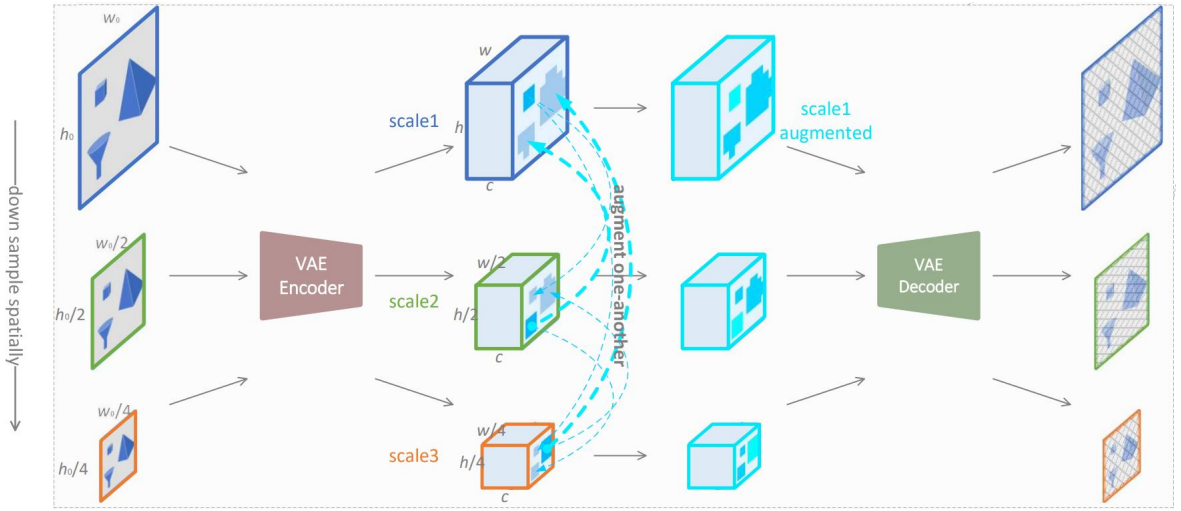
- However, existing VAE guidance does not explicitly address that
  - objects can vary in pixel sizes, i.e., multi-scale patterns,
  - while models typically excel at specific pattern scales.
- Multi-scale techniques are well explored in detection and segmentation tasks,
  - but we are the first to realize this in OCL setting,
  - with novel design specifically for VAE models.

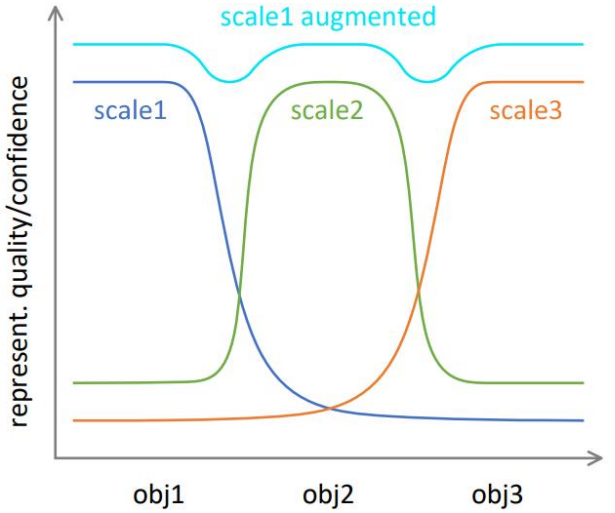




Method

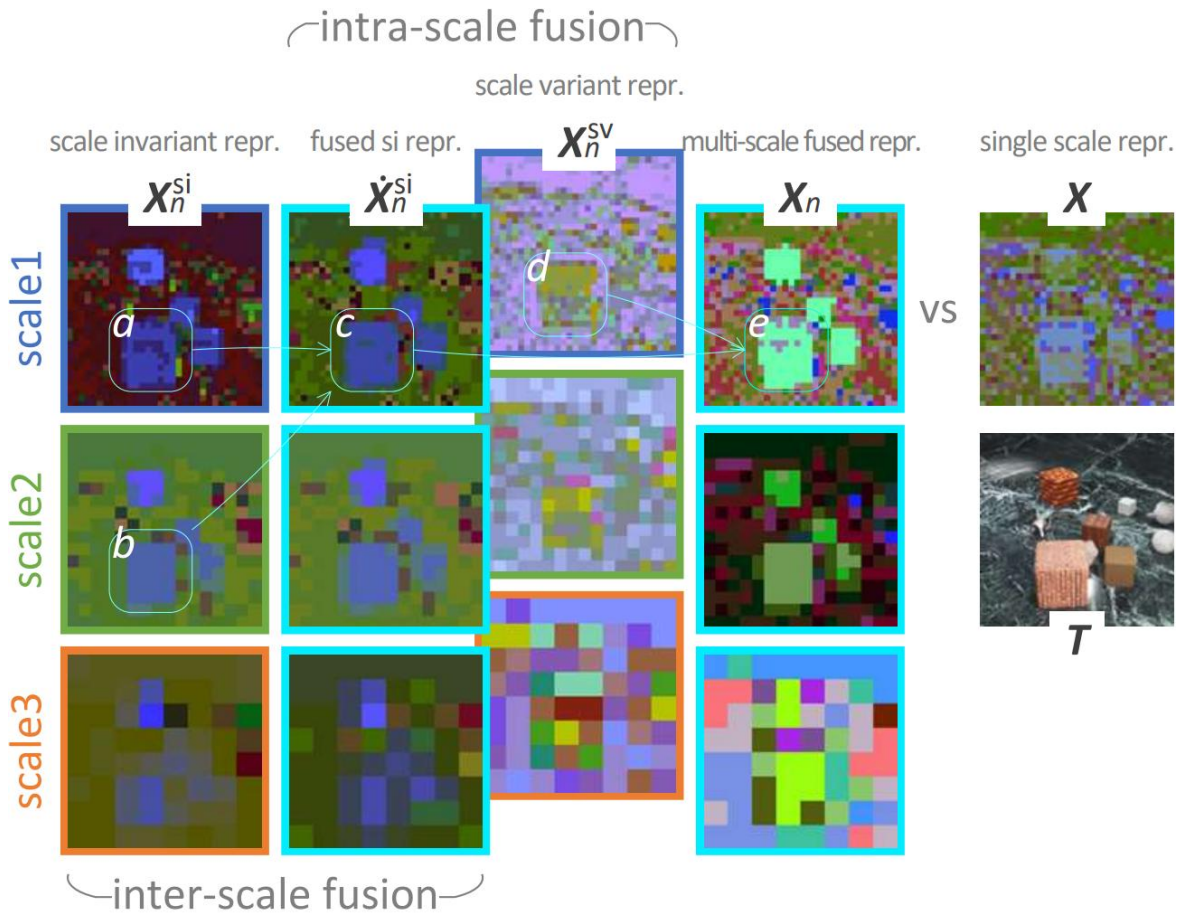
- We propose Multi-Scale Fusion (MSF) to enhance VAE guidance for OCL training.
- To ensure objects of all sizes fall within VAE's comfort zone,
  - we adopt the image pyramid,
    - which produces intermediate representations at multiple scales;
- To foster scale-invariance/variance in object super-pixels,
  - we devise inter/intra-scale fusion,
    - which augments low quality object super-pixels of one scale with corresponding high-quality super pixels from another scale.





Demonstration

- Our MSF shows interpretable visual patterns in its fusions.



Comparison

- MSF vs VAR (NeurIPS 2024 Best Paper) and SPAE (NeurIPS 2023 Spotlight)
  - Both re-sample the same input feature map into different scales,
  - no inter-scale and intra-scale fusion
  - MSF re-samples the input image to produce different scales,
  - with novel inter-scale and intra-scale fusion

SPAE	Multi-Scale VAE Represent.	resample feature map	no fusion
VAR			
MSF		resample input image	inter-/intra-scale fusion

Experiment

- Our MSF applies to different methods and boosts their performance.

	ARI	ARI_{fg}	mIoU	mBO
		ClevrTex		
$SLATE_r$	20.56	68.14	34.11	34.57
+SysBind@2	23.27	71.27	35.63	36.62
+GDR@$g2$	34.46	73.38	37.42	36.69
+MSF	32.70	80.70	40.61	41.48
		COCO		
$SLATE_r$	24.18	24.54	21.37	21.76
+SysBind@2	25.71	24.97	21.46	22.01
+GDR@$g2$	30.37	29.95	23.47	22.98
+MSF	30.95	30.47	23.33	23.85
		VOC		
$SLATE_r$	11.64	15.65	15.64	14.99
+SysBind@2	11.75	16.04	15.73	15.01
+GDR@$g2$	13.20	17.49	16.46	16.65
+MSF	12.17	16.54	16.74	16.69

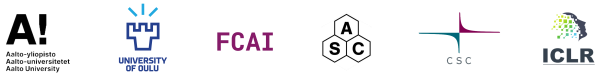
Table 1: Transformer-based *image* OCL performance on synthetic and real-world datasets.

	ARI	ARI_{fg}	mIoU	mBO
		MOVi-C		
$STEVE_c$	52.20	31.16	16.74	19.05
+SysBind@2	51.87	34.64	17.36	19.12
+GDR@$g2$	60.14	35.79	20.01	21.95
+MSF	60.94	36.22	20.33	22.74
		MOVi-D		
$STEVE_c$	35.71	50.24	19.10	20.88
+SysBind@2	35.34	52.16	19.46	21.53
+GDR@$g2$	40.37	52.47	20.42	22.64
+MSF	43.20	55.64	21.21	23.14
		MOVi-E		
$STEVE_c$	28.00	52.06	18.78	20.48
+SysBind@2	28.47	55.46	18.95	20.50
+GDR@$g2$	34.17	53.21	19.47	20.76
+MSF	36.70	54.28	20.39	22.37

Table 2: Transformer-based *video* OCL on synthetic datasets.

	ARI	ARI_{fg}	mIoU	mBO
		ClevrTex		
$SlotDiffuz_r$	64.21	26.50	31.51	32.44
+GDR@$g2$	69.21	37.83	34.74	34.03
+MSF	71.47	38.08	35.06	35.67
DINOSAUR	60.74	45.75	30.48	32.56
		COCO		
$SlotDiffuz_r$	36.54	35.67	22.08	22.75
+GDR@$g2$	37.68	36.33	22.73	22.25
+MSF	37.63	36.99	22.32	22.87
DINOSAUR	33.24	33.35	22.01	21.93
		VOC		
$SlotDiffuz_r$	16.97	14.33	15.71	16.02
+GDR@$g2$	18.20	15.59	16.11	17.04
+MSF	19.40	15.69	16.37	16.76
DINOSAUR	16.00	18.48	15.94	16.37

Table 3: Diffusion-based *image* OCL on synthetic and real-world datasets. * Subscript "r" and "c" stand for random and condition query initialization, respectively.


A!
Aalto-yliopisto
Aalto-universitetet
Aalto University
UNIVERSITY OF OULU
FCAI
A
S
C
CSC
ICLR

Thanks!