Smoothing Slot Attention *Iterations* and *Recurrences*

Rongzhen Zhao 1
Wenyan Yang 1
Juho Kannala 2,3
Joni Pajarinen 1
1 Aalto EEA; 2 Aalto CS; 3 Oulu CMVS

more about my research and me

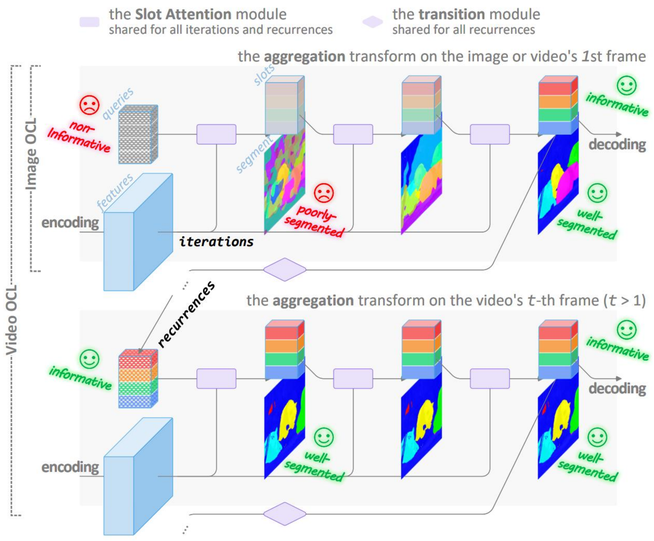
paper source code model checkpoints training logs

What is Object-Centric Learning (OCL):

- discover objects in images or videos without supervision;
- represent them as corresponding feature vectors, slots.

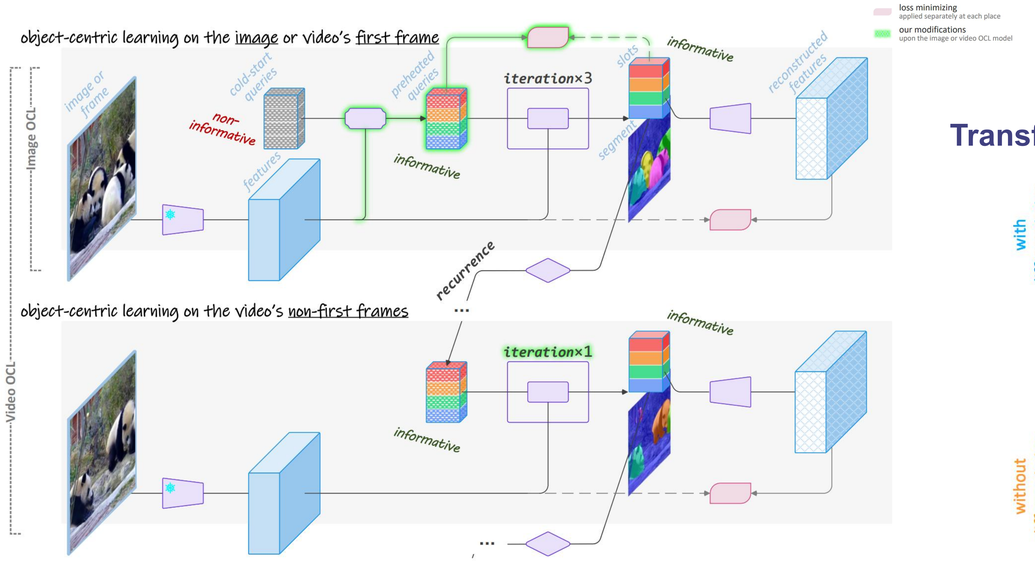
How is an OCL model trained:

- aggregate input's dense features into slots;
- reconstruct the input from such slots;
- minimize the reconstruction error.

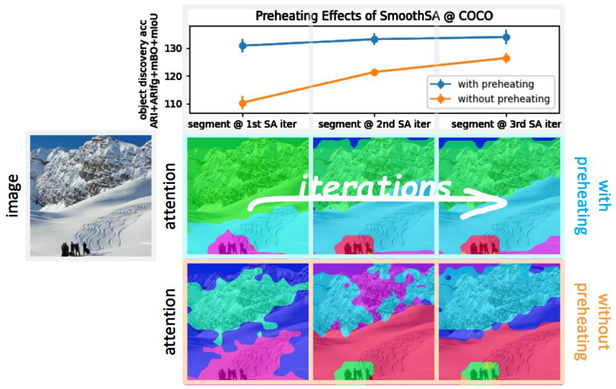


We address these issues with our SmoothSA:

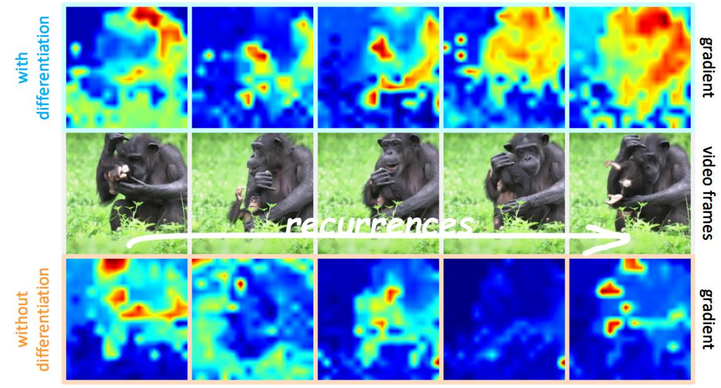
- To smooth SA iterations on image or video's first frame, we *preheat cold-start* queries with rich input-feature information, by a tiny module self-distilled inside OCL;
- To smooth SA recurrences across video's first and non-first frames, we *differentiate* the *homogeneous* aggregation transforms by using full and single iterations respectively



Query preheating smooths SA iterations:



Transform differentiation smooths SA recurrences:



Slot Attention (SA) in mainstream OCL:

- Image features can be aggregated into object-level representations by SA *iteratively* refining cold-start query slots.
- For video, such aggregation proceeds by SA *recurrently* shared across frames, with queries cold-started on the first frame while transitioned from the previous frame's slots thereafter.

Issues in current SA iteration and recurrence:

- *Cold-start* queries lack sample-specific cues thus hindering precise aggregation on image or video's first frame;
- Non-first frame queries are already sample-specific thus requiring aggregation transforms *different* from the first frame

SmoothSA boosts performance effectively:

	ClevrTex #slot=11				COCO #slot=7				VOC #slot=6			
	ARI	ARI_{fg}	mBO	mIoU	ARI	ARI_{fg}	mBO	mIoU	ARI	ARI_{fg}	mBO	mIoU
SLATE	$17.4_{\pm 2.9}$	$87.4_{\pm 1.7}$	$44.5_{\pm 2.2}$	$43.3_{\pm 2.4}$	$17.5_{\pm 0.6}$	$28.8_{\pm 0.3}$	$26.8_{\pm 0.3}$	$25.4_{\pm 0.3}$	$18.6_{\pm 0.1}$	$26.2_{\pm 0.8}$	$37.2_{\pm 0.5}$	$36.1_{\pm 0.4}$
DINOSAUR	$50.7_{\pm 24.1}$	$89.4_{\pm 0.3}$	$53.3_{\pm 5.0}$	$52.8_{\pm 5.2}$	$18.2_{\pm 1.0}$	$37.0_{\pm 1.2}$	$28.3_{\pm 0.5}$	$26.9_{\pm 0.5}$	$21.5_{\pm 0.7}$	$36.2_{\pm 1.3}$	$40.6_{\pm 0.6}$	$39.7_{\pm 0.6}$
SlotDiffusion	$66.1_{\pm 1.3}$	$82.7_{\pm 1.6}$	$54.3_{\pm 0.5}$	$53.4_{\pm 0.8}$	$17.7_{\pm 0.5}$	$29.0_{\pm 0.1}$	$27.0_{\pm 0.4}$	$25.6_{\pm 0.4}$	$17.0_{\pm 1.2}$	$21.7_{\pm 1.8}$	$35.2_{\pm 0.9}$	$34.0_{\pm 1.0}$
SPOT	$25.6_{\pm 1.4}$	$77.1_{\pm 0.5}$	$48.2_{\pm 0.6}$	$46.3_{\pm 0.7}$	$23.7_{\pm 0.5}$	$40.4_{\pm 0.5}$	$30.9_{\pm 0.2}$	$29.3_{\pm 0.2}$	$24.5_{\pm 0.3}$	$31.0_{\pm 0.8}$	$40.1_{\pm 0.2}$	$38.6_{\pm 0.3}$
DIAS	$80.9_{\pm 0.3}$	$79.1_{\pm 0.3}$	$63.3_{\pm 0.1}$	$61.9_{\pm 0.0}$	$25.6_{\pm 0.1}$	$41.2_{\pm 0.3}$	$31.7_{\pm 0.1}$	$30.2_{\pm 0.1}$	$30.9_{\pm 0.5}$	$33.5_{\pm 0.7}$	$43.4_{\pm 0.5}$	$42.4_{\pm 0.5}$
SmoothSA^i	$76.8_{\pm 0.7}$	$82.2_{\pm 1.7}$	$60.6_{\pm 0.4}$	$58.9_{\pm 0.6}$	$29.3_{\pm 1.0}$	$41.3_{\pm 1.2}$	$33.4_{\pm 0.2}$	$31.8_{\pm 0.2}$	$35.0_{\pm 0.5}$	$33.6_{\pm 1.2}$	$45.2_{\pm 0.3}$	$43.9_{\pm 0.3}$

Table 1. Object discovery on images. Input resolution is 224×224; DINO2 ViT-S/14 is for encoding.

	MOVi-C #slot=11, conditional				MOVi-D #slot=21, conditional				YTVIS-HQ #slot=7			
	ARI	ARI_{fg}	mBO	mIoU	ARI	ARI_{fg}	mBO	mIoU	ARI	ARI_{fg}	mBO	mIoU
STEVE	–	–	–	–	$17.5_{\pm 0.6}$	$28.8_{\pm 0.3}$	$26.8_{\pm 0.3}$	$25.4_{\pm 0.3}$	–	–	–	–
VideoSAUR	$41.9_{\pm 1.1}$	$53.3_{\pm 2.1}$	$16.1_{\pm 0.4}$	$14.8_{\pm 0.4}$	$22.5_{\pm 5.0}$	$40.0_{\pm 20.1}$	$11.6_{\pm 6.6}$	$10.8_{\pm 6.1}$	$33.8_{\pm 0.7}$	$49.2_{\pm 0.5}$	$29.9_{\pm 0.4}$	$29.7_{\pm 0.4}$
SlotContrast	$64.6_{\pm 9.4}$	$59.9_{\pm 5.3}$	$27.7_{\pm 3.0}$	$25.8_{\pm 2.9}$	$45.3_{\pm 4.1}$	$63.9_{\pm 0.2}$	$26.7_{\pm 1.0}$	$25.1_{\pm 1.0}$	$37.2_{\pm 0.6}$	$49.4_{\pm 1.1}$	$33.0_{\pm 0.2}$	$32.8_{\pm 0.1}$
RandSF.Q	$65.4_{\pm 10.7}$	$67.4_{\pm 2.1}$	$29.2_{\pm 3.8}$	$26.8_{\pm 3.7}$	$41.6_{\pm 3.7}$	$77.5_{\pm 1.0}$	$27.4_{\pm 1.0}$	$25.6_{\pm 1.0}$	$40.1_{\pm 0.4}$	$58.0_{\pm 1.0}$	$37.6_{\pm 0.4}$	$37.2_{\pm 0.4}$
SmoothSA^v	$50.9_{\pm 1.6}$	$69.0_{\pm 0.3}$	$31.7_{\pm 0.8}$	$30.2_{\pm 0.8}$	$43.8_{\pm 1.5}$	$70.5_{\pm 0.7}$	$31.4_{\pm 0.4}$	$30.2_{\pm 0.4}$	$42.4_{\pm 0.8}$	$63.0_{\pm 3.4}$	$38.9_{\pm 0.7}$	$38.3_{\pm 0.6}$
update 20260425 below												
	MOVi-C #slot=11, conditional				MOVi-E #slot=24, conditional				YTVIS-2022 #slot=7			
	ARI	ARI_{fg}	mBO	mIoU	ARI	ARI_{fg}	mBO	mIoU	ARI	ARI_{fg}	mBO	mIoU
VideoSAUR	$41.9_{\pm 1.1}$	$53.3_{\pm 2.1}$	$16.1_{\pm 0.4}$	$14.8_{\pm 0.4}$	$17.4_{\pm 2.5}$	$34.6_{\pm 20.7}$	$8.3_{\pm 4.9}$	$7.5_{\pm 4.3}$	$33.4_{\pm 0.8}$	$48.2_{\pm 0.7}$	$27.2_{\pm 0.3}$	$26.8_{\pm 0.3}$
SlotContrast	$64.6_{\pm 9.4}$	$59.9_{\pm 5.3}$	$27.7_{\pm 3.0}$	$25.8_{\pm 2.9}$	$29.9_{\pm 4.9}$	$70.6_{\pm 3.8}$	$20.7_{\pm 1.4}$	$19.3_{\pm 1.2}$	$35.2_{\pm 0.8}$	$51.4_{\pm 0.7}$	$29.7_{\pm 0.5}$	$29.3_{\pm 0.6}$
RandSF.Q	$65.4_{\pm 10.7}$	$67.4_{\pm 2.1}$	$29.2_{\pm 3.8}$	$26.8_{\pm 3.7}$	$30.5_{\pm 1.2}$	$82.1_{\pm 3.1}$	$23.0_{\pm 1.2}$	$21.6_{\pm 1.4}$	$37.9_{\pm 1.3}$	$51.8_{\pm 1.2}$	$32.2_{\pm 1.8}$	$31.5_{\pm 1.8}$
SmoothSA^v	$50.9_{\pm 1.6}$	$69.0_{\pm 0.3}$	$31.7_{\pm 0.8}$	$30.2_{\pm 0.8}$	$36.7_{\pm 0.6}$	$73.6_{\pm 0.6}$	$28.6_{\pm 0.1}$	$27.4_{\pm 0.1}$	$42.0_{\pm 0.6}$	$59.0_{\pm 2.1}$	$36.0_{\pm 0.5}$	$34.9_{\pm 0.6}$

Table 2. Object discovery on videos. Input resolution is 224×224; DINO2 ViT-S/14 is for encoding.

		class top1	top3	bbox IoU	#match
			COCO #slot=7		
SPOT	+MLP	$88.7_{\pm 0.3}$	$97.2_{\pm 0.1}$	$48.6_{\pm 0.1}$	$5061_{\pm 30}$
SmoothSA^i	+MLP	$89.3_{\pm 0.5}$	$97.3_{\pm 0.2}$	$49.5_{\pm 1.1}$	$5210_{\pm 223}$
			YTVIS-HQ #slot=7		
SlotContrast	+MLP	$85.8_{\pm 0.3}$	$95.8_{\pm 0.4}$	$51.5_{\pm 0.5}$	$9249_{\pm 41}$
SmoothSA^v	+MLP	$90.4_{\pm 0.2}$	$97.6_{\pm 0.1}$	$42.6_{\pm 1.4}$	$8957_{\pm 34}$
update 20260425 below					
		class top1	top3	bbox IoU	#match
			YTVIS-2022 #slot=7		
SlotContrast	+MLP	$87.1_{\pm 0.2}$	$96.4_{\pm 0.1}$	$48.2_{\pm 0.3}$	$19943_{\pm 156}$
SmoothSA^v	+MLP	$91.0_{\pm 0.2}$	$97.6_{\pm 0.0}$	$39.2_{\pm 1.0}$	$18864_{\pm 200}$

Table 3. Object recognition on images and videos.

			GQA #slot=7
			accuracy %
SPOT	+ Aloe		$52.3_{\pm 2.8}$
SmoothSA^i	+ Aloe		$56.7_{\pm 1.9}$
			CLEVRER #slot=7
			per option % / per question %
SlotContrast	+ Aloe		$97.2_{\pm 1.1}$ / $95.6_{\pm 0.9}$
SmoothSA^v	+ Aloe		$98.7_{\pm 0.4}$ / $96.9_{\pm 0.6}$

Table 4. Visual question answering on images and videos.