From Vicious to Virtuous Cycles: Synergistic Representation Learning for Unsupervised Video Object-Centric Learning

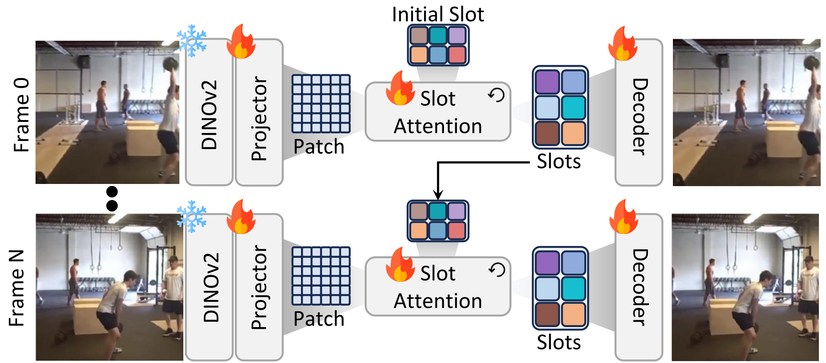
Hyun Seok Seong*, WonJun Moon* and Jae-Pil Heo (*: equal contribution)
Sungkyunkwan University

Code

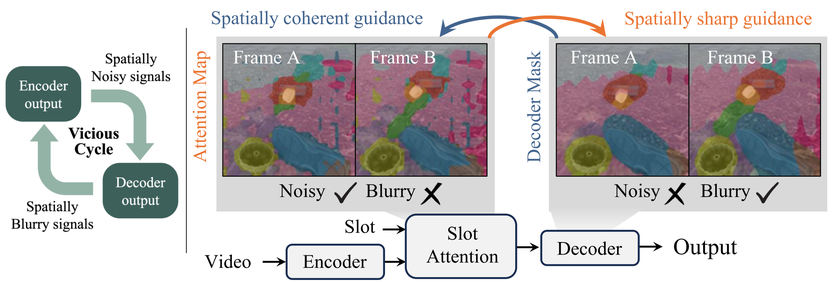
Video Object-Centric Learning

Object-Centric Learning is to group images into object entities without supervision.

- Slot-attention is widely employed technique to implement object-centric learning.
- For temporal domain, slot-attention typically recurrently used.



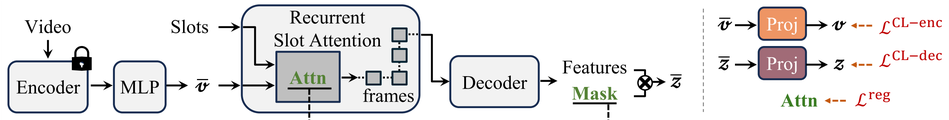
Motivation

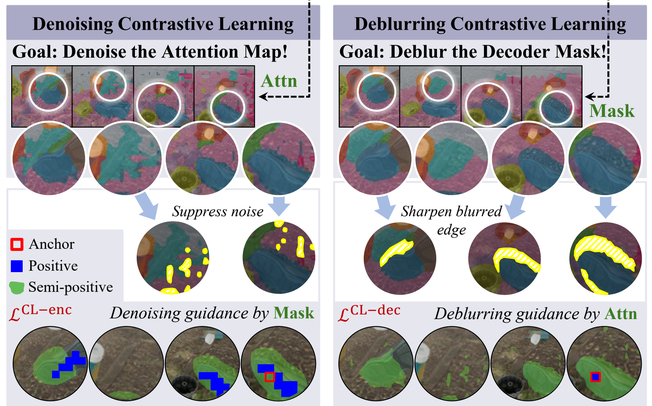


- Both spatial maps are optimized with global reconstruction loss.

However, **Noisy Encoder** renders the decoder's reconstruction task ill-posed.
In turn, **Blurry Decoder** tends to generate more blurry object maps as the safest strategy to minimize the reconstruction loss is to average over the possibilities.
Consequently, blurry outputs yield **uninformative gradients**, which lacks the structural detail needed to refine the encoder's sharp but noisy features.

Synergistic Representation Learning





Denoising Contrastive Learning (Refining encoder with decoder)

Goal: Suppress background noise in encoder attention maps.
Mechanism: Utilizes spatially coherent decoder masks as guidance.
Effect: Pulls noisy attention toward clean attention, denoising the slot representations.

■ Positive : Top-K similar patches from the backbone ■ Semi-positive : Patches assigned to the same slot in the decoder mask.

Deblurring Contrastive Learning (Refining decoder with encoder)

Goal: Sharpen the blurry boundaries of decoder object masks.
Mechanism: Leverages sharp encoder attention as guidance.
Effect: Refines structural details of decoder mask.

■ Positive : Corresponding patch in the reconstruction target. ■ Semi-positive : Patches assigned to the same slot in the attention mask.

Slot Regularization

Goal: Prevent redundant slot allocation and over-fragmentation (splitting object into multi-parts).
Mechanism: Identify spatially redundant slots & reinitialize slot w/ high-entropy.
Effect: Encourages stable & distinct object discovery,
which alleviates the risk of propagating noisy semi-positive targets in SRL.

Experimental Results

- Experiment results on MOVi-C, MOVi-E, and YouTube-VIS 2021 datasets.

Method	MOVi-C		MOVi-E		YouTube-VIS	
	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑
SAVi (Kipf et al., 2021)	22.2	13.6	42.8	16.0	-	-
STEVE (Singh et al., 2022)	36.1	26.5	50.6	26.6	15.0	19.1
VideoSAUR (Zadaianchuk et al., 2023)	64.8	**38.9**	73.9	**35.6**	28.9	26.3
VideoSAURv2 (Manasyan et al., 2025)	–	–	77.1	34.4	31.2	29.7
SlotContrast (Manasyan et al., 2025)	69.3	32.7	82.9	29.2	38.0	33.7
SlotContrast† (Manasyan et al., 2025)	70.4	31.7	80.9	28.2	36.2	32.9
SRL (Ours)	**74.3**	34.5	**81.9**	29.3	**42.9**	**35.6**

- Downstream applications on object dynamics prediction.

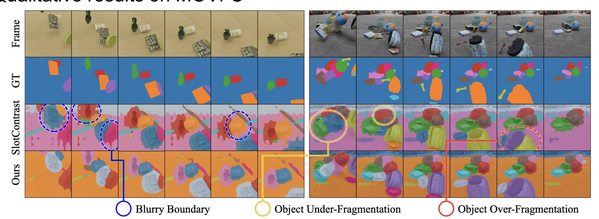
Method	MOVi-C		MOVi-E		YouTube-VIS	
	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑
Reconstruction + SF	50.7	25.9	**70.6**	24.3	27.4	28.9
SlotContrast + SF	63.8	26.1	70.5	**24.9**	29.2	29.6
SRL (Ours) + SF	**68.9**	**27.4**	70.4	**24.9**	**32.2**	**30.0**

Deblur $\mathcal{L}^{\text{CL-dec}}$	Denoise $\mathcal{L}^{\text{CL-enc}}$	Reg $\mathcal{L}^{\text{reg}}$	MOVi-C FG-ARI↑	MOVi-C mBO↑
			70.8	31.4
✓			70.0	33.2
	✓		72.2	31.2
✓	✓		70.7	35.1
✓		✓	73.0	33.5
	✓	✓	74.2	33.2
✓	✓	✓	74.3	34.5

- Ablation Study

While denoising and deblurring are effective, slot regularization unlocks their full potential.
By ensuring slots capture distinct objects, it prevents the model from inadvertently sharpening false boundaries between over-fragmented object parts.

- Qualitative results on MOVi-C



- Qualitative results on Youtube-VIS 2021