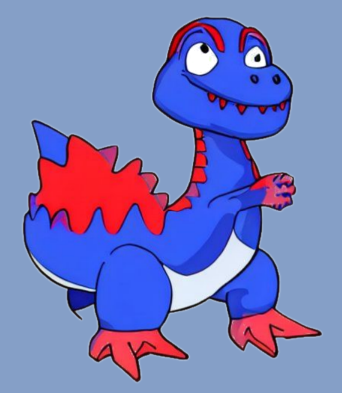
Object-Centric Learning for Real-World Videos by Predicting Temporal Feature Similarities

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Andrii Zadaianchuk* Maximilian Seitzer* Georg Martius



Summary

- We present VideoSAUR (**Video** **S**lot **A**ttention **U**sing temporal feature similaRity): *the first video object-centric method that scales to unconstrained real-world datasets covering diverse domains.*
- We greatly **outperform previous state-of-the-art methods** on challenging synthetic datasets.
- VideoSAUR is the first video-based object-centric method to scale to the YouTube-VIS dataset.

Prior Work: Recurrent Slot Attention

- Slot Attention-based models follow an encoder-decoder framework with a set-vectored bottleneck.
- The Slot Attention module groups input features into slots via iterative, competitive attention steps.
- Recurrent Slot Attention initializes the slots using slots of the previous frame.

```
# inputs: feature maps + position embedding
def recurrent_slot_attention(inputs, slots_prev, t):
  # Slot recurrence: init random or from prev. slots
  if t == 0:
    slots = random_normal(mean, std)
  else:
    slots = predictor(slots_prev)

  # N iterations of slot attention
  for n in range(N):
    scores = dot(k(inputs), q(slots))
    weights = softmax(scores, axis="slots")
    updates = weighted_mean(weights, v(inputs))
    slots = gru(slots, updates)
    slots = slots + mlp(slots)
  return slots
```

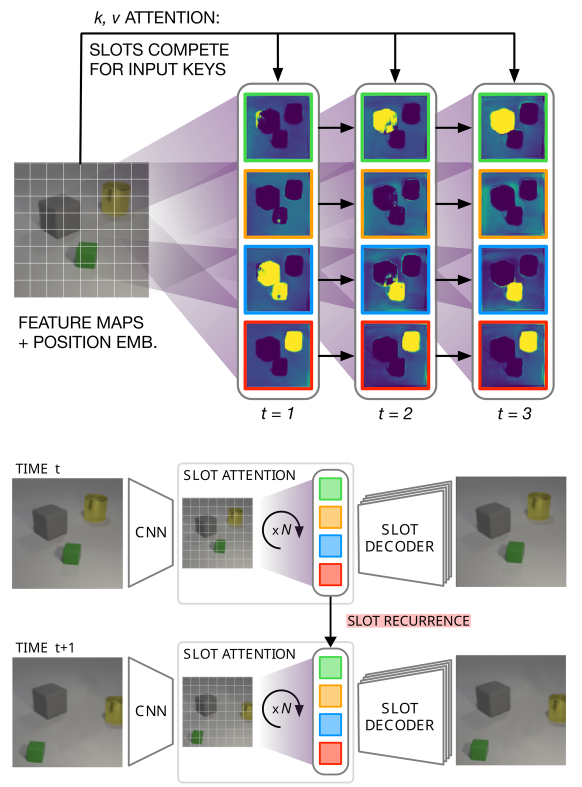
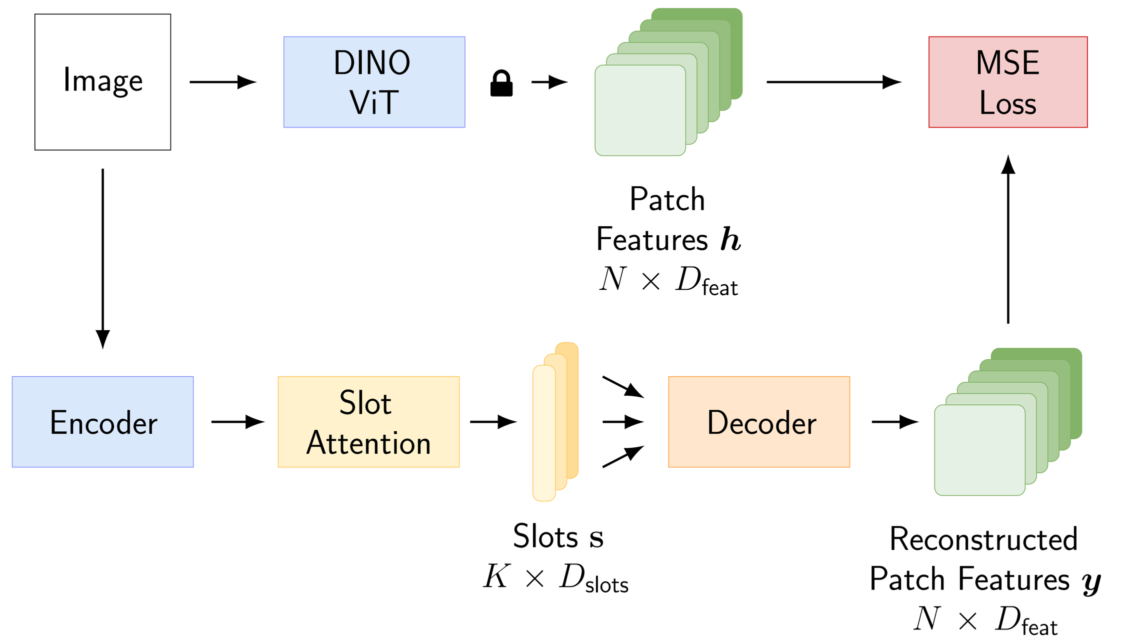


Figure provided by courtesy of the authors of [3].

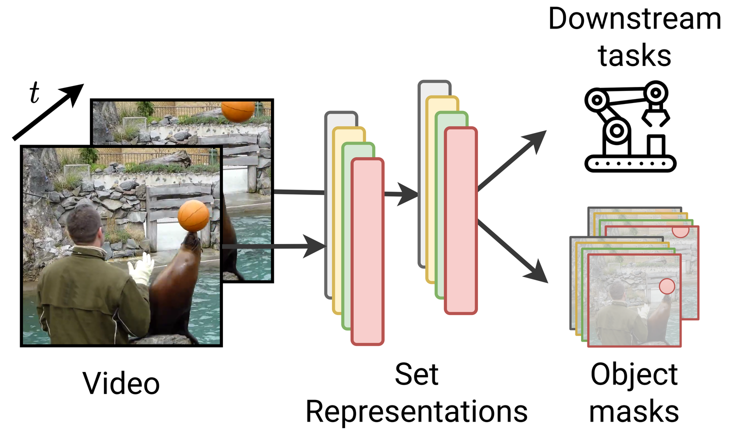
Prior Work: DINOSAUR

- Our previous work DINOSAUR (ICLR'23, [4]) was the first object-centric model scaling to *real-world image data* (e.g. PASCAL VOC, COCO).
- DINOSAUR utilizes pre-trained, highly semantic self-supervised features (e.g. DINO [1]) with a feature reconstruction objective.



Video Object-Centric Learning

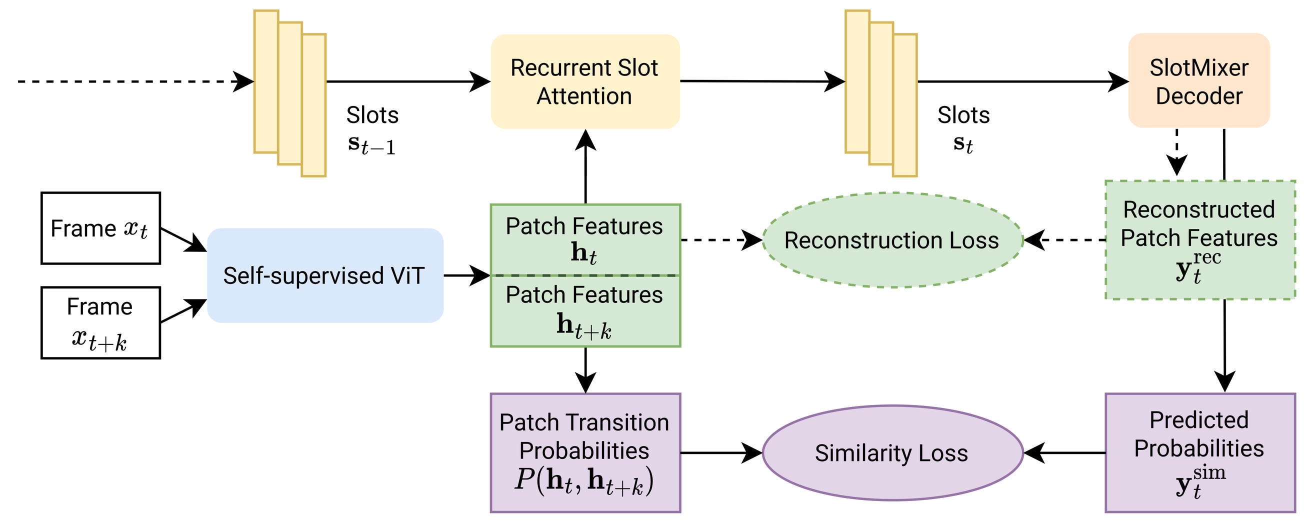
- Represent video frames as a set of vectors.
- Maintain *consistency* of the representation in time.
- Produce localization masks for each representation.

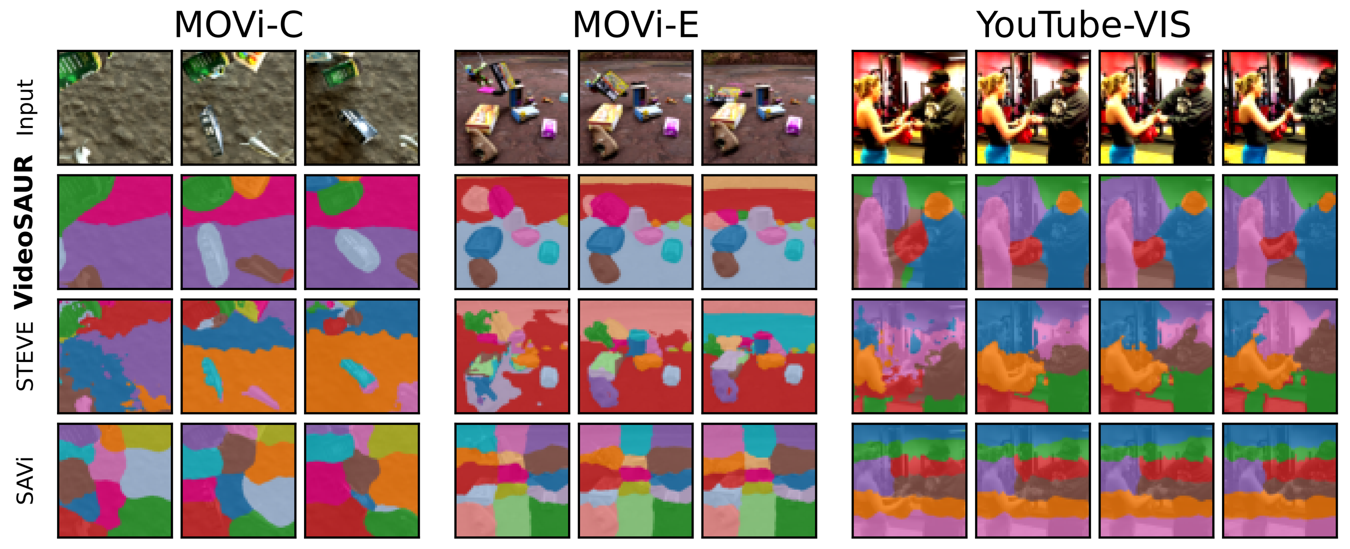


Method

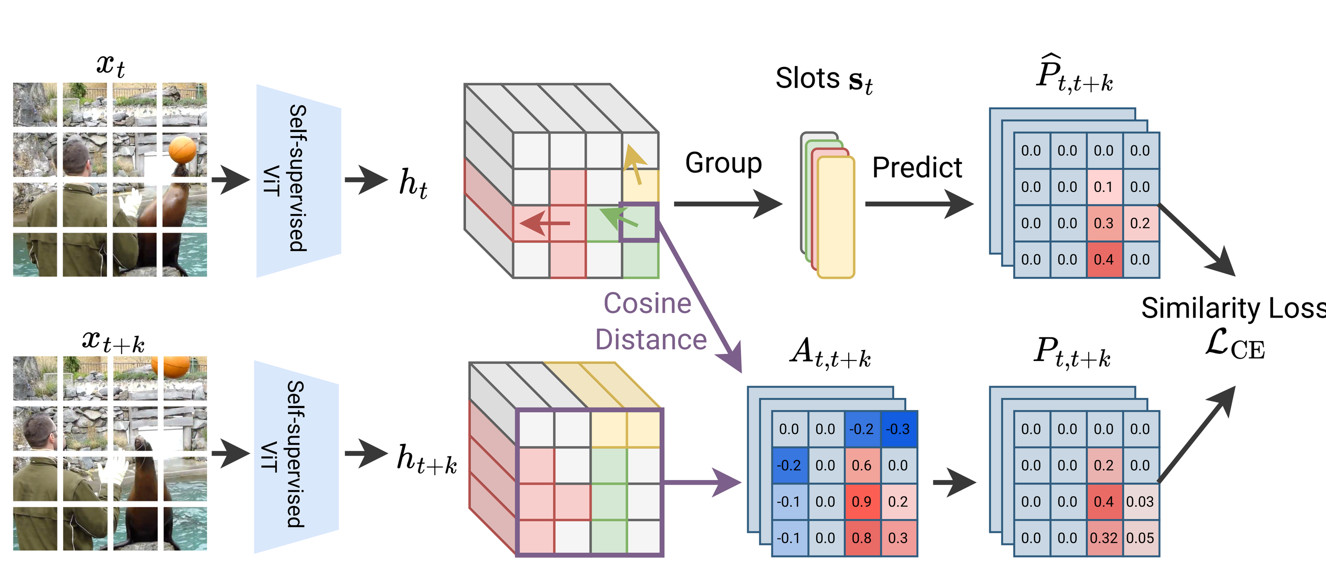
- We combine Recurrent Slot Attention [2] with DINOSAUR [4] and add a *temporal similarity loss* that exploits temporal and semantic correlations for object grouping.
- The temporal similarity loss incentivizes grouping patches
  - with similar motion (similar to optical flow prediction).
  - with similar semantics (useful e.g. for static objects).
- For efficient video decoding, we integrate the SlotMixer decoder that scales well with the number of slots.
- Loss function: temporal similarity $\mathcal{L}^{\text{sim}}$, optionally reconstruction loss $\mathcal{L}^{\text{rec}}$

$$\mathcal{L} = \sum_{t=1}^{T-k} \mathcal{L}^{\text{sim}}(\boldsymbol{P}_{t,t+k}, \boldsymbol{y}_t^{\text{sim}}) + \alpha \mathcal{L}^{\text{rec}}(\boldsymbol{h}_t, \boldsymbol{y}_t^{\text{rec}}). \tag{1}$$





Temporal Similarity Loss



- Given L patch features $\boldsymbol{h} \in \mathbb{R}^{L \times D}$ from times t and $t+k$, we compute the affinity matrix $\boldsymbol{A}_{t,t+k} \in [-1,1]^{L \times L}$:

$$\boldsymbol{A}_{t,t+k} = \frac{\boldsymbol{h}_t}{\|\boldsymbol{h}_t\|} \cdot \left(\frac{\boldsymbol{h}_{t+k}}{\|\boldsymbol{h}_{t+k}\|}\right)^{\top}, \tag{2}$$

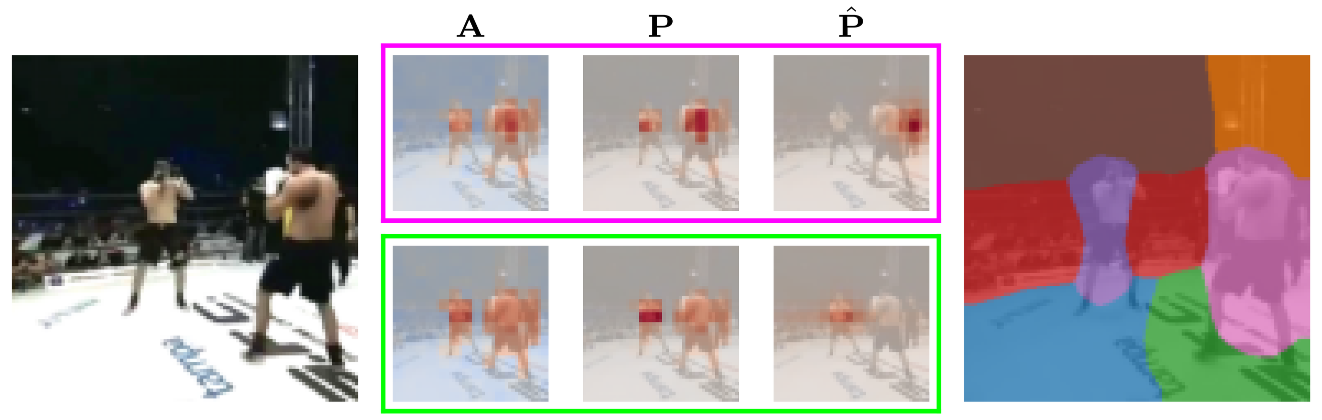
and normalize it to a transition probability matrix $\boldsymbol{P}_{t,t+k}$:

$$\boldsymbol{P}_{t,t+k} = \text{softmax}\left(\frac{\boldsymbol{A}_{t,t+k}}{\tau}, \text{ axis} = t+k\right). \tag{3}$$

- Model *predicts the transition probabilities* $\boldsymbol{y}_t^{\text{sim}} = \widehat{\boldsymbol{P}}_{t,t+k}$ for each patch:

$$\mathcal{L}^{\text{sim}} = \text{CE}(\boldsymbol{P}_{t,t+k}; \widehat{\boldsymbol{P}}_{t,t+k}). \tag{4}$$

- Example affinity matrices $\boldsymbol{A}$, probabilities $\boldsymbol{P}$ and predictions $\widehat{\boldsymbol{P}}$:



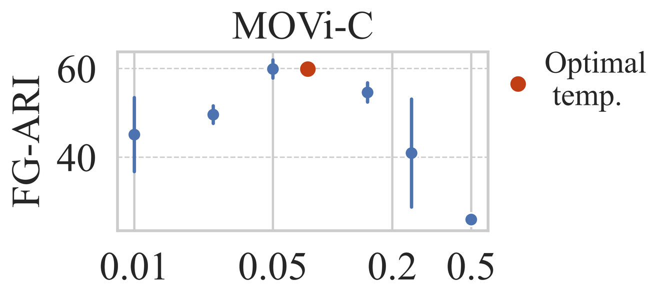
Comparison to Object-Centric Methods

- We compare with SotA video object-centric methods on challenging synthetic datasets (MOVi) and real-world datasets (YouTube-VIS).

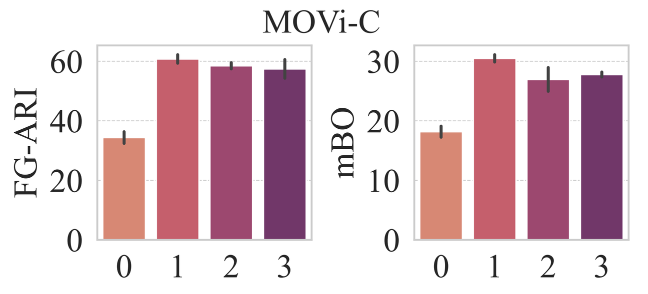
	MOVi-C		MOVi-E		YT-VIS	
	FG-ARI	mBO	FG-ARI	mBO	FG-ARI	mBO
Block Pattern	24.2	11.1	36.0	16.5	24	14.9
SAVi	22.2	13.6	42.8	16.0	11.1	12.7
STEVE	36.1	26.5	50.6	26.6	20.0	20.9
VideoSAUR	**64.8**	**38.9**	**73.9**	**35.6**	**39.5**	**29.1**

Analysis of Similarity Loss Parameters

- Temperature τ controls prediction task reliance on motion/semantics:

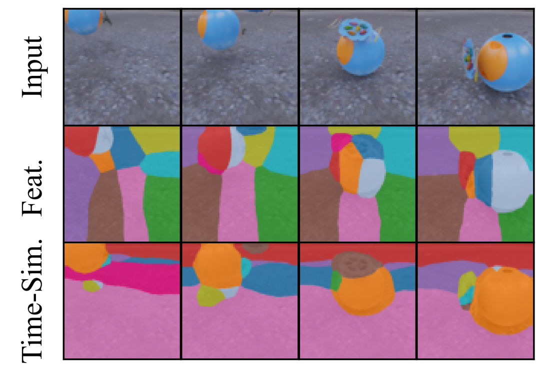


- Time-offset k affects self-supervised task difficulty:



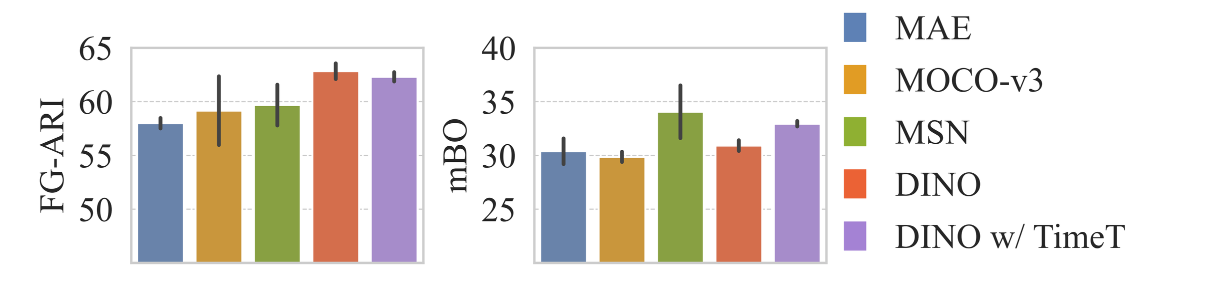
Loss Ablations

Feat. Rec.	Next Feat.	Frame Pred.	Temp. Sim.	Optical Flow	MOVi-C FG-ARI	MOVi-C mBO	YT-VIS FG-ARI	YT-VIS mBO
✓					40.2	23.5	35.4	26.7
				✓	48.9	—	—	—
✓	✓				47.2	24.7	37.9	27.3
			✓		**60.8**	**30.5**	26.2	**29.1**
✓			✓		60.7	30.3	**39.5**	**29.1**



Choice of Self-Supervised Features

- VideoSAUR performs well with different ImageNet self-supervised features...



- ...but also with features pre-trained directly on the target domain (MOVi):

	MOVi-C		MOVi-E	
	FG-ARI	mBO	FG-ARI	mBO
MAE, ImageNet pretraining	58.0	30.4	72.8	27.1
MAE, MOVi-E pretraining	59.8	27.5	70.6	23.3

Surprising! ImageNet's object-centric bias is apparently *not needed*.

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging Properties in Self-Supervised Vision Transformers. *ICCV*, 2021.

[2] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff. Conditional Object-centric Learning from Video. In *ICLR*, 2022.

[3] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf. Object-Centric Learning with Slot Attention. In *NeurIPS*, 2020.

[4] M. Seitzer, M. Horn, A. Zadaianchuk, D. Zietlow, T. Xiao, C.-J. Simon-Gabriel, T. He, Z. Zhang, B. Schölkopf, T. Brox, and F. Locatello. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023.

Webpage