Semantics Meets Temporal Correspondence: Self-supervised Object-centric Learning in Videos

Rui Qian, Shuangrui Ding, Xian Liu, Dahua Lin

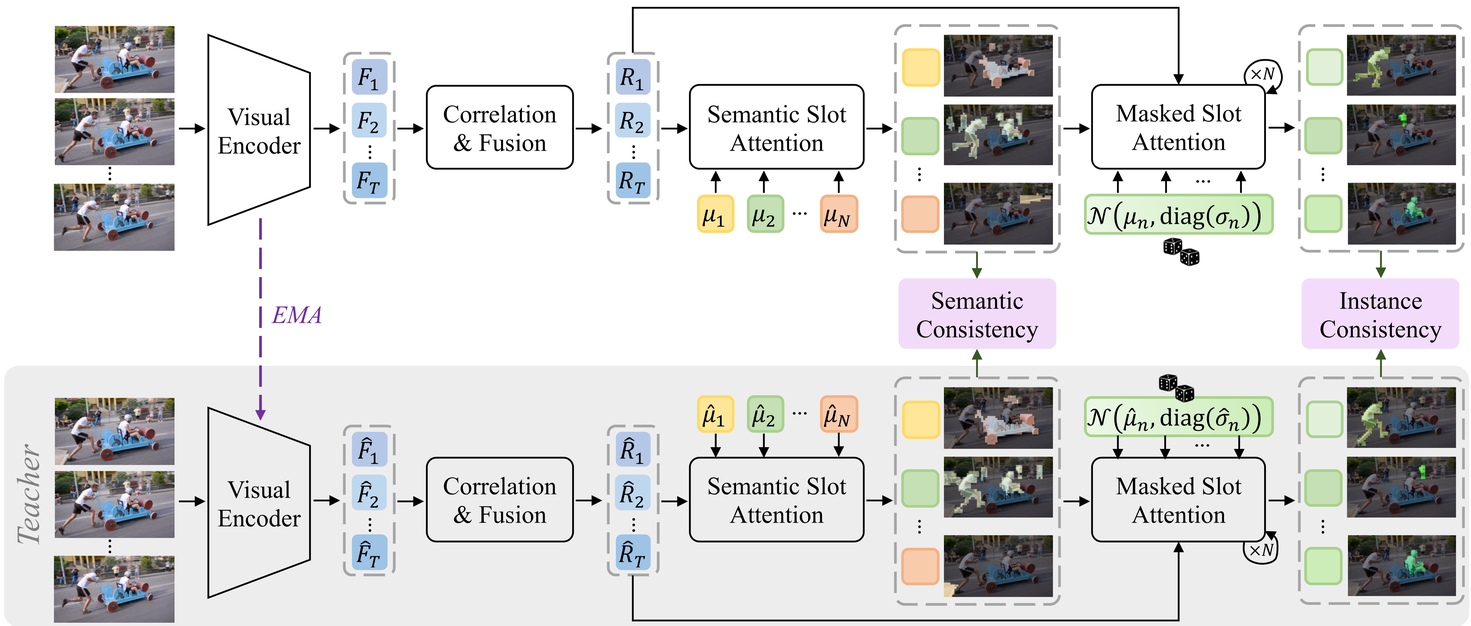
ICCV23 PARIS

Abstract:

- **Task:** Object-centric video representation learning, including unsupervised video object segmentation and spatio-temporal correspondence.
- **Motivation:** Jointly utilize high-level semantics and low-level correspondence to parse the object structures. There exist consistent semantic patterns for similar objects across scenes, while the temporal correspondence cues vary with specific scenes.
- **Our solution:** We propose a two-stage semantic-aware slot attention for semantic decomposition and instance discrimination. We distill temporally coherent object-centric representations in a fully self-supervised manner.

Method:

- Fuse the frame-wise feature and simple temporal correlation map to carry on rich semantics and temporal correspondence cues.
- Formulate a set of learnable Gaussian distributions consisting of mean and deviation vectors. The former represents potential semantic centers, the latter introduces random perturbation to capture temporal correspondence patterns.
- Develop two-stage semantic-aware slot attention, which first uses mean vectors as slot initialization for semantic decomposition, then performs random sampling around each semantic center to identify instances within the semantic area.



- **Semantic Consistency:** First determine dense patch correspondence across time with optimal transport, then align the semantic distributions.

$$\mathcal{L}_{sem} = -\sum_{\substack{t,j=1 \\ t\neq j}}^{T} \sum_{u,v=1}^{HW} \sum_{n=1}^{N} \pi^{*}_{tj}[u,v]\widehat{\mathcal{M}}_j[n,v]\log\mathcal{M}_t[n,u]$$

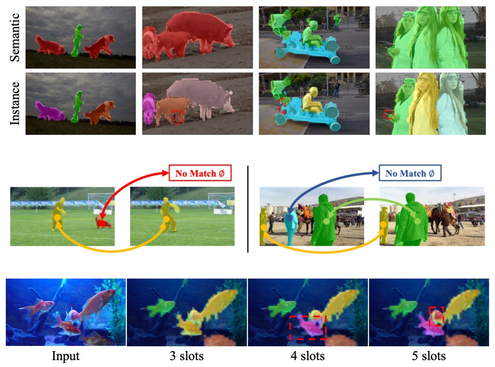
- **Instance Consistency:** First filter out non-existing semantic centers and invalid instance slots, then encourage temporally coherent object representations.

$$\mathcal{L}_{obj} = -\sum_{\substack{t,j=1 \\ t\neq j}}^{T} \sum_{n=1}^{N} \sum_{p=1}^{P} \mathcal{I}_t[n,p] \left\{ \hat{\mathcal{I}}_j[n,\varepsilon(p)] \left|\left| \mathcal{O}_t[n,p] - \hat{\mathcal{O}}_j[n,\varepsilon(p)] \right|\right|_2 + \sum_{\substack{q=1 \\ q\neq\varepsilon(p)}}^{P} \mathrm{relu}\left(\lambda - \left|\left| \mathcal{O}_t[n,p] - \hat{\mathcal{O}}_j[n,q] \right|\right|_2\right) \right\}$$

Experiment:

- **Unsupervised video object discovery:** Single object segmentation on DAVIS-2016, SegTrack-v2, FBMS-59; Multiple object segmentation on DAVIS-2017-Unsupervised.

Model	RGB	Flow	DAVIS	ST-v2	FBMS
CIS [89]	✓	✓	71.5	62.5	63.5
AMD [57]	✓	✗	57.8	57.0	47.5
DINO [12]	✓	✗	52.3	46.5	50.3
SIMO [53]	✗	✓	67.8	62.0	-
MG [88]	✗	✓	68.3	58.6	53.1
OCLR [84]	✗	✓	**72.1**	67.6	65.4
GWM [16]	✓	✓	71.2	69.0	66.9
SMTC	✓	✗	71.8	**69.3**	**68.4**
SMTC†	✓	✗	70.8	68.4	66.5



Model	Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
DINOSAUR [72]	ViT-S/16	21.4	19.2	23.7
SMTC	ViT-S/16	**40.5**	**36.4**	**44.6**
SMTC	ResNet-50	39.0	35.5	42.6
RVOS [76]	ResNet-101	41.2	36.8	45.7
ProReduce [56]	ResNet-101	68.3	65.0	71.6

- **Dense spatio-temporal correspondence:** Semi-supervised video object segmentation on DAVIS-2017; Pose tracking on JHMDB; Human part tracking on VIP.

		DAVIS			JHMDB		VIP
Model	Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	PCK@0.1	PCK@0.2	mIoU
Supervised [40]	ResNet-50	66.0	63.7	68.4	59.2	78.3	39.5
MoCo [38]	ResNet-50	65.4	63.2	67.6	60.4	79.3	36.1
VFS [86]	ResNet-50	68.9	66.5	71.3	60.9	80.7	43.2
DINO [12]	ViT-S/16	61.8	60.2	63.4	45.4	75.2	37.9
TimeCycle [81]	ResNet-50	40.7	41.9	39.4	57.7	78.5	28.9
UVC [55]	ResNet-50	56.3	54.5	58.1	56.5	76.6	34.2
MAST [51]	ResNet-18	65.5	63.3	67.6	-	-	-
CRW [44]	ResNet-18	67.6	64.8	70.2	58.8	80.3	37.6
SFC [43]	ResNet-18+ResNet-50	71.2	68.3	74.0	61.9	83.0	38.4
SMTC	ViT-S/16	67.6	64.1	71.2	53.2	79.6	**39.2**
SMTC	ResNet-50	**73.0**	**69.4**	**76.6**	**62.5**	**84.1**	38.8

Conclusion:

- Unify semantic discrimination and temporal correspondence for object-centric video representation learning.
- Achieve fully self-supervised video object instance identification with semantic structure.
- Efficiently transfer object knowledge to more general video understanding tasks.