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

Rui Qian[1] Shuangrui Ding[1] Xian Liu[1] Dahua Lin[1,2*]
[1]The Chinese University of Hong Kong [2]Shanghai Artificial Intelligence Laboratory
{qr021,ds023,lx021,dhlin}@ie.cuhk.edu.hk

1. More Implementation Details

Optimal Transport Solution. For dense semantic distribution alignment, we formulate it as an optimal transport problem as:

$$\begin{aligned}\min_{\pi_{tj}} & \sum_{u=1}^{HW}\sum_{v=1}^{HW} -\hat{C}_{tj}[u,v]\pi_{tj}[u,v] \\ \text{s.t.} & \sum_{v=1}^{HW}\pi_{tj}[\cdot,v] = \frac{1}{HW}\mathbf{1}^{HW}, \\ & \sum_{u=1}^{HW}\pi_{tj}[u,\cdot] = \frac{1}{HW}\mathbf{1}^{HW}, \\ & \pi_{tj}[u,v] \geq 0 \quad u,v \in \{1,2,...,HW\},\end{aligned} \tag{1}$$

where t, j is respectively temporal index, u, v is spatial index, $-\hat{C}_{ij}$ denotes the transport cost, π_{tj} is the transportation strategy, and the marginal distributions on source and target are set to uniform distribution without requiring prior in default. Inspired by [3, 1], we employ Sinkhorn-Knopp algorithm [4] to solve this problem. In specific, we aim to solve the following objective with regularization:

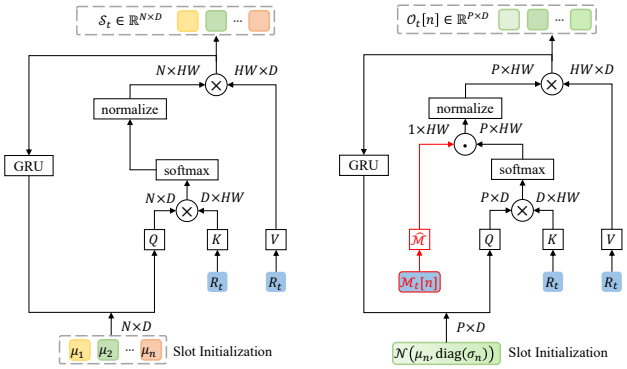
$$\min_{\pi_{tj}} \sum_{u,v} -\hat{C}_{ij}[u,v]\pi_{ij}[u,v] + \epsilon \sum_{u,v} \pi_{ij}[u,v]\log \pi_{ij}[u,v]. \tag{2}$$

And the optimal solution can be written as

$$\pi^*_{ij} = \text{Diag}(x)\exp(-\frac{c_{ij}}{\epsilon})\text{Diag}(y), \tag{3}$$

where $x \in \mathbb{R}^{HW}$ and $y \in \mathbb{R}^{HW}$ are renormalization vectors calculated by iterative Sinkhorn-Knopp algorithm. We set the hyper-parameter $\epsilon = 0.05$, and use 3 iterations in default.

*Corresponding author. Email: dhlin@ie.cuhk.edu.hk.



(a) Semantic Decomposition. (b) Instance Discrimination.

Figure 1. We compare the details of two slot attention stages. For semantic decomposition, we initialize the slots with learnable mean vectors, then follow standard slot attention iterations to generate semantic masks and semantic center representations as indicated in Fig. 1(a). For instance discrimination, we randomly sample vectors from each learnable Gaussian distribution as slot initialization for each semantics, then perform masked slot attention with obtained semantic mask as reference and generate instance segmentation and representations. We use red line to indicate the difference between the masked slot attention and standard slot attention in Fig. 1(b).

Semantic-aware Masked Slot Attention Formulation. We present the details of two slot attention stages in our semantic-aware masked slot attention, i.e., semantic-aware slot attention stage for semantic decomposition, masked slot attention stage for instance discrimination. We compare the two slot attention stages in Fig. 1.

For the first stage semantic-aware slot attention designed for semantic decomposition, we use the learnable mean vectors $\boldsymbol{\mu} = \{\mu_1, \mu_2, ..., \mu_N\} \in \mathbb{R}^{N\times D}$ as slot initialization, each representing a potential semantic center. Note that the `softmax` and `normalize` in Fig. 1(a) respectively denotes the softmax operation on slot dimension N, and $\mathcal{L}_1$ normalization on spatial dimension HW. The iterative at-

Encoder	Dist.	IoU	$\mathcal{J}\&\mathcal{F}$
ViT-S/16	Uniform	71.8	40.5
	CAAM	72.1	40.8
	Attention	73.3	41.3
ResNet-50	Uniform	66.4	39.0
	CAAM	68.1	39.8

Table 1. Ablation studies on the marginal distribution formulation in optimal transport. We compare default uniform distribution, normalized class-agnostic activation map [2], and the attention score between `cls` token and spatial feature tokens from the last layer of ViT encoder. We report the results on single object benchmark DAVIS-2016 and multiple object benchmark DAVIS-2017-Unsupervised.

tention calculation and slot update are the same as the standard slot attention in [5], and finally output semantic segmentation masks of size $\mathbb{R}^{N\times HW}$, semantic center representations of size $\mathbb{R}^{N\times D}$.

For the second stage masked slot attention designed for instance discrimination, we run this stage on N semantics in parallel, i.e., first discriminate instances of the same semantics then aggregate the results of all semantics. In specific, for n-th semantics, we randomly sample vectors from the learnable Gaussian distribution $\mathcal{N}(\mu_n, \text{diag}(\sigma_n))$ as slot initialization to represent P potential instances of n-th semantics. Next, in each iteration, the difference with standard slot attention is denoted in red in Fig. 1(b). We use the semantic masks computed in the first stage as reference to only preserve the visual contents related to n-th semantics and filter out unrelated spatial areas. We use this masked slot attention weight to aggregate `value` features and update the slots. Finally, for each semantics, it outputs instance masks of size $\mathbb{R}^{P\times HW}$ as well as instance representations of size $\mathbb{R}^{P\times D}$. Aggregating all N semantics, we obtain $N \times P$ instances in total with individual segmentation masks of size $\mathbb{R}^{HW}$ and representations of size $\mathbb{R}^{D}$.

2. More Experimental Results

Marginal Distribution in Optimal Transport. In Table 1, we delve into the detailed formulation of the optimal transport, Eq 1, to determine patch correspondence. In default, we use the uniform distribution as the marginal distribution following [3, 4]. We compare using two forms of semantic prior: class-agnostic activation map (CAAM) [2] as well as the attention between `cls` token and other spatial feature tokens on ViT backbone as the marginal distribution to assign larger importance weights to foreground areas. The performance improvement on both single and multiple object discovery demonstrates that such semantic guidance enables the model to lay more emphasis on object areas and enhances object-centric representations.

Number P	IoU	$\mathcal{J}\&\mathcal{F}$
1	66.4	24.5
2	69.3	35.4
3	71.1	39.9
4	71.8	40.5
5	71.5	40.4

Table 2. Ablation studies on the number of potential instances P. We vary the number from 1 to 5, and report the results on single object benchmark DAVIS-2016 and multiple object benchmark DAVIS-2017-Unsupervised.

Number of Potential Instances. In Table 2, we compare different number of potential instances P in the second slot stage for instance discrimination. When $P = 1$, our formulation degenerates into semantic slot attention which only decomposes semantics without discriminating instances. Hence, it performs much worse on multiple object discovery benchmark but achieves comparable results on single object cases. When we increase P, the performance slightly improves then maintains stable. It demonstrates that our method is generally robust to this hyperparameter.

And another interesting phenomenon of slot attention is that in inference, it is feasible to sample different number of slots to generalize to various scenes with distinct number of objects [5]. And our method also maintains this attribute as validated in Fig. 2. We show the qualitative comparison of using different number of slots in training and inference. We observe that when the number of objects is larger than the number P defined in training, it is practical to sample more slots in inference to discriminate distinct instances. Specifically, for the model trained with $P = 3$, we respectively sample $P = 3, 4, 5$ slots in inference on the gold-fish sequence where there are five different fishes in the scene. When the sampled number is lower than the object number, there are instances grouped together which are discovered by the same slot. When we increase the number of sampled slots in inference, our model gradually improves the granularity of instance identification, with the red box highlighting the changing area.

Temporal Correspondence Sampling. Recall that in Sec. 3.1 in main submission, for timestamp t, we randomly sample one temporal index $j \neq t$ to calculate the temporal correspondence map C_{tj}. Here, we conduct ablation studies on sampling different number of temporal indexes in Table 3. Specifically, we design a baseline where for timestamp t, we sample t-th frame itself to calculate the correspondence map. In this way, the calculated C_{tt} is equivalent to self-correlation with no temporal cues, and the performance significant drops especially on multiple object discovery. It is consistent with our intuition that temporal

Figure 2. Different number of instance slots P in inference. We use $P = 3$ in training, and visualize the results of $P = 3, 4, 5$ in inference on the gold-fish sequence. The red box denotes the major changing area with various P in inference.

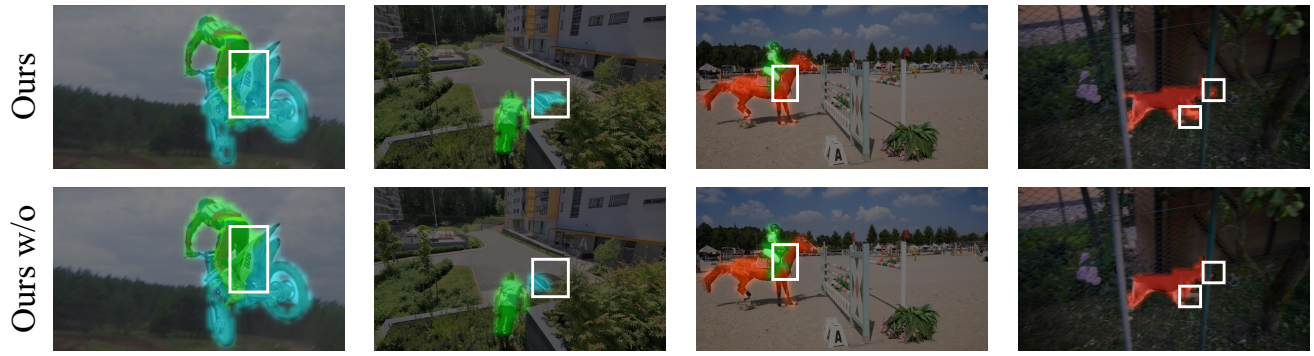


Figure 3. Comparison of semantic decomposition map. We respectively present our method with and without the second instance-level slot attention stage in training in the first and second row. We use white boxes to highlight the difference area.

Sampling	Number	IoU	$\mathcal{J}\&\mathcal{F}$
Self	1	67.1	24.3
Other	1	71.8	40.5
Other	2	71.7	40.8
Other	3	71.8	40.7

Table 3. Ablation studies on the sampling strategy in temporal correspondence calculation. 'Self' denotes for each frame, we calculate the self-correlation as the correspondence map. 'Other' means sampling other frames to calculate correspondence. We report the results on single object benchmark DAVIS-2016 and multiple object benchmark DAVIS-2017-Unsupervised.

correspondence cues contribute to identifying different instances of the same semantics. And for sampling different frames to calculate correspondence, we compare sampling different number of frames within the input clip. When we sample multiple frames, we respectively calculate the correspondence of each pair and make an average as the final correspondence representation. For example, we sample two indexes $i \neq t$ and $j \neq t$, we take the average $\frac{C_{ti}+C_{tj}}{2}$ to fuse with semantic feature map. We observe that sampling more frames could slightly facilitate multiple object discovery due to more abundant temporal information. But generally, sampling one temporal index is sufficient to provide the temporal correspondence cue to supplement semantic feature and assist instance identification.

3. More Qualitative Results

Semantic Decomposition Map. We compare the semantic decomposition map with two different training formulations in Fig. 3. For more straightforward comparison, we visualize the soft masks without binarization to exclude the impact of threshold. In the first row, we present the semantic decomposition results of our method. In the second row, we show the results of our method with only first slot attention stage in training, i.e., no instance-level discrimination and alignment is considered. We observe that though the results are both generated by the first semantic slot attention stage, the instance-level understanding in training could facilitate semantic understanding. For example, the instance-level understanding helps to generate more precise and clear borders of the person and motorbike, and improves the object part awareness when dealing with the occluded car and dog.

Instance Discrimination Map. We also present more visualization results on the final instance-level segmentation maps in Fig. 4. In the first row, we show the instance discovery results of our method. In the second row, we present the results of our method without masked feature aggregation in the second slot attention stage. From the comparison, it echos with our motivation that the semantic mask as a prior reference could enforce the instance slots to concentrate on specific semantic areas and improve the instance discrimination results. For example, the masked feature ag-

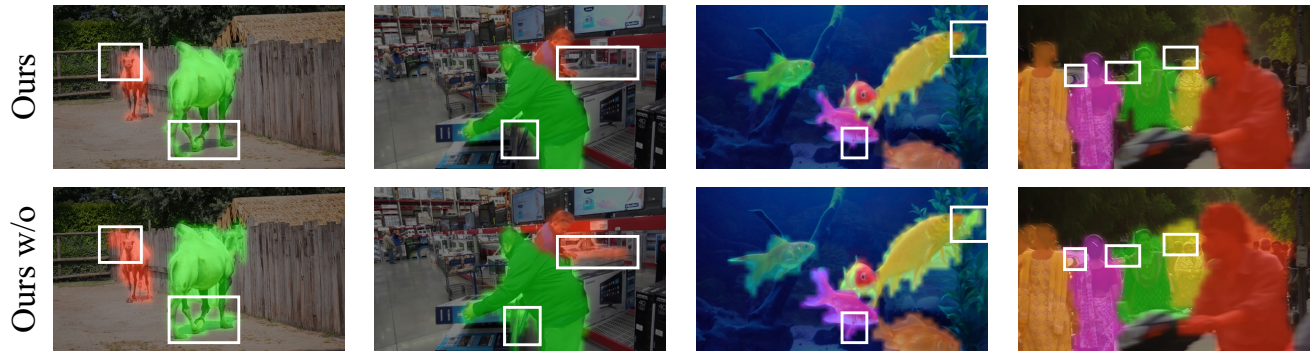


Figure 4. Comparison of instance identification map. We respectively present our method with and without the masked aggregation in the slot attention stage in the first and second row. We use white boxes to highlight the difference area.

gregation suppresses the distracting background areas of the fence, and leads to clearer borders that discriminate difference person or fish instances.

Figure 5. The iterative evolving of slot attention. We present an example on the dog, and show 3 iterations of the slot attention map.

Iterative Evolving of Slot Attention Map. We also give an example on iterative evolving of the learned slots in Fig. 5. We observe that the slots first attend to the parts with most salient attributes, e.g., legs, then gradually expand to the whole object. It demonstrates the necessity of using iterative slot attention for comprehensive object discovery, preventing the model from only localizing object parts.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

[2] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10451–10459, 2020.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[5] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.