Betrayed by Attention: A Simple yet Effective Approach for Self-supervised Video Object Segmentation

Shuangrui Ding[1]*, Rui Qian[1]*,
Haohang Xu[2], Dahua Lin[1,3,4]†, and Hongkai Xiong[2]

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[4] Centre for Perceptual and Interactive Intelligence (CPII), Hong Kong, China
`{ds023,qr021}@ie.cuhk.edu.hk`

Abstract. In this paper, we propose a simple yet effective approach for self-supervised video object segmentation (VOS). Previous self-supervised VOS techniques majorly resort to auxiliary modalities or utilize iterative slot attention to assist in object discovery, which restricts their general applicability. To deal with these challenges, we develop a simplified architecture that capitalizes on the emerging objectness from DINO-pretrained Transformers, bypassing the need for additional modalities or slot attention. Our key insight is that the inherent structural dependencies present in DINO-pretrained Transformers can be leveraged to establish robust spatio-temporal correspondences in videos. Furthermore, simple clustering on this correspondence cue is sufficient to yield competitive segmentation results. Specifically, we first introduce a single spatio-temporal Transformer block to process the frame-wise DINO features and establish spatio-temporal dependencies in the form of self-attention. Subsequently, utilizing these attention maps, we implement hierarchical clustering to generate object segmentation masks. To train the spatio-temporal block in a fully self-supervised manner, we employ semantic and dynamic motion consistency coupled with entropy normalization. Our method demonstrates state-of-the-art performance across three multi-object video segmentation tasks. Specifically, we achieve over 5 points of improvement in terms of FG-ARI on complex real-world DAVIS-17-Unsupervised and YouTube-VIS-19 compared to the previous best result. The code and checkpoint are released at `https://github.com/shvdiwnkozbw/SSL-UVOS`.

Keywords: Video Object Segmentation · Self-supervised Learning

---

* Equal Contribution

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

<div align="center">(a) Video Frame   (b) DINO Attention   (c) Random Initialized Spatio-temporal Attention   (d) Learned Spatio-temporal Attention</div>
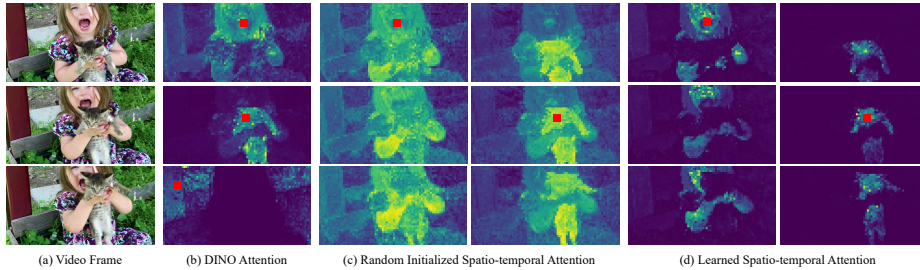
Fig. 1: Attention leaks the object's position! We visualize the self-attention maps of different queries (■ prompt) from the video sequence (a). The frame-wise DINO attention maps (b) highlight image regions corresponding to the queried object. A randomly initialized spatio-temporal Transformer block on top of DINO produces noisy spatio-temporal attention maps (c) that coarsely track objects over time. Our method diminishes noise in the learned spatio-temporal attention maps (d) which derive temporally coherent object segmentation.

1   Introduction

Representing the visual scene with objects as the basic elements has been long acknowledged as a core cognitive capability of an intelligent agent. In the realm of computer vision, Video Object Segmentation (VOS) tasks require a model to segment and track specified objects within a video sequence, striving to emulate this foundational function. This task holds significant importance in various real-world vision systems, including but not limited to, autonomous driving [18] and surveillance security [43]. However, traditional methods for VOS [6, 44] typically follow a fully supervised paradigm and entail substantial costs for obtaining pixel-level per-frame annotations. This factor significantly limits their feasibility for large-scale applications. Thus, the researchers turn their attention to self-supervised learning [8, 9, 11, 12, 23, 38, 47, 50, 66], a more efficient approach that leverages unlabeled data.

Despite its promising potential, self-supervised VOS poses significant challenges. Firstly, previous methods tend to incorporate additional signals such as optical flow [36, 68, 71], or depth cues [16, 29], to furnish object-related clues. However, these auxiliary signals within videos can often be difficult to extract and unreliable when it comes to multi-object scenarios, thus limiting their availability. Secondly, most existing models [1, 48, 74] adopt slot attention [39] to parse the frame into objects. Nevertheless, the requirement for a predefined quantity of learnable slot queries limits its adaptability to in-the-wild data, especially when the number of objects is unknown. To address these limitations, we recall the recent self-supervised learning technique DINO [7, 42], which demonstrates emerging objectness in the attention maps of pretrained Vision Transformer (ViT) [14]. As depicted in Fig. 1(b), DINO attention maps encode spatial dependencies between diverse patches. These dependencies present different patterns for various objects, providing abundant cues for object segmentation.

This phenomenon motivates us with an intriguing possibility: Can we utilize the object-aware attributes of DINO to learn robust spatio-temporal dependencies and thereby produce coherent object segmentations in videos?

In this paper, we delve into the potential of this idea with a simple architecture. Based on DINO pretrained ViT, we introduce a single spatio-temporal Transformer block to process frame-wise DINO features and calculate spatio-temporal attention maps. Remarkably, these attention maps in Fig. 1(c) showcase the ability to coarsely discriminate and track object parts even with random initialization. Based on this observation, we can roughly estimate the patches belonging to the same object over time. Then, we encourage both semantic and motion alignment among patches belonging to the same object while distinguishing those from different objects in a fully self-supervised manner. The training significantly diminishes noise in the learned spatio-temporal dependencies as displayed in Fig. 1(d). In inference, we directly apply Hierarchical Clustering [28] to these spatio-temporal attention maps to derive object segmentation masks within a sequence. Interestingly, this naive clustering method yields unexpectedly competitive results. Our method greatly outperforms slot attention-based methods on various datasets, e.g., **9** points improvement over SOLV [1] on DAVIS-17, **12** points improvement over SMTC [48] on YTVIS-19. We argue that clustering on the whole spatio-temporal attention maps, instead of frame-by-frame feature clustering as in slot attention ensures the temporal coherence of produced segmentations and guarantees the object's permanence.

In summary, our proposed method **BA** (short for **B**etrayed by **A**ttention) offers three distinct advantages: (1) BA solely requires RGB frames and operates entirely under self-supervision without the need for external modality. This characteristic renders our method highly accessible and adaptable in real-world scenarios; (2) The architecture is remarkably simple and systematically efficient. We only introduce a single learnable Transformer block and harness a parameter-free clustering algorithm to discover the objects. It facilitates easy generalization to multi-object scenes, without the constraints imposed by slot attention; (3) Our method achieves state-of-the-art results across a bunch of unsupervised video multi-object segmentation benchmarks, such as MOVi-E [19], DAVIS-17-Unsupervised [46], and YouTube-VIS-19 [69]. Notably, we surpass the previous state-of-the-art method, TimeT [52], by 6.6 points on YTVIS-19 and by 6.1 points on DAVIS-17 in terms of the FG-ARI metric.

2   Related Work

Video Object Segmentation aims to segment objects coherently in a video sequence [2,6,15,17,20,21,25,27,33,44]. In video object segmentation (VOS), there are two prevalent protocols for evaluating the learned models: Semi-supervised VOS and Unsupervised VOS. In semi-supervised VOS, the algorithm is provided with the object masks in the first frame, and required to track them in subsequent frames. In contrast, unsupervised VOS aims to identify and segment salient objects from the background without any specific reference. In this paper, we fo-

cus on the more challenging unsupervised VOS setting without using any kind of manual annotations in either training or inference. Recently, there emerges a line of self-supervised algorithms for unsupervised VOS [10, 29, 34, 58, 64, 68, 70–72]. CIS [71] facilitates fully unsupervised motion segmentation, discarding object mask supervision during training. By formulating a min-max game of mutual information, the generator is motivated to create segmentations that effectively distinguish foreground objects from the background. AMD [37] minimizes the warping synthesis error to train appearance and motion pathways without any supervision. GWM [10] uses a single RGB image as input with optical flow acting as supervision to highlight moving areas. MG [68] and OCLR [65] solely leverage optical flow as input to generate object-centric layered representations, with each layer indicating a potential object. Despite the impressive performance of state-of-the-art self-supervised VOS methods [1, 48, 52] in discovering single objects, they still face limitations when it comes to solving multi-object discovery tasks. In contrast, by leveraging the rich object cues in attention maps, our method can discern multiple objects in real-world scenarios.

Self-supervised Spatio-temporal Correspondence learning usually leverages free temporal supervision signals in videos to learn representations that facilitate accurate pixel or object tracking across space and time [4, 7, 13, 24, 26, 30, 31, 35, 57, 59, 67]. Vondrick et al. [57] employ the natural temporal color coherence to train a colorization model on grayscale videos, thereby establishing fine-grained correspondence between current and future frames. CRW [26] presents a self-supervised learning approach for dense correspondence in raw videos. It uses space-time graph-based random walks and cycle consistency to implicitly supervise chains of comparisons. Hu et al. [24] independently learn semantics and temporal correspondence from two pathways and fuse them at a later stage. UME [32] respectively designs short-term appearance and long-term semantic consistency to learn generalizable correspondence. Taking a step further by integrating high-level semantics with low-level temporal correlation, SMTC [48] develops two-stage slot attention to establish dense correspondence with more emphasis on foreground objects. In this work, we explicitly initialize the spatio-temporal correspondence on top of the DINO pretrained Transformer. This initialization effectively guides the learning process with a self-supervised pixel-level consistency.

Object-aware DINO Features have shown efficacy in object localization tasks, particularly in image domain [5, 22, 41, 54, 60, 63, 73, 76]. DINO's original paper [7] reveals that the learned representations and attention maps carry substantial object cues that can be harnessed for object discovery. The groundbreaking work, LOST [54], utilizes these features from DINO to perform object segmentation by constructing a graph wherein objects are segmented using the inverse degrees of nodes. TokenCut [62] employs DINO features for applying the Normalized Cut algorithm [53], thereby obtaining foreground segments in an image. Furthermore, CutLER [60] and MOST [51] can localize multiple objects without any supervision based on the DINO features. Motivated by the successful application of DINO features in image settings, recent works have expanded

their use to video segmentation tasks [1,48,52,61,62]. SMTC [48] utilizes the rich semantics and correspondence cues in the DINO features to provide a reliable reference for object decomposition in consecutive frames. TimeT [52] introduces a feature forwarding process to propagate dense DINO features across time and establish temporal consistency. Most of the existing works focus on leveraging DINO feature vectors for optimization and adaptation to videos. While in our work, we lay more emphasis on the self-attention maps. We extend the frame-wise DINO attention maps to spatio-temporal maps, which explicitly encode space-time structural dependencies and help coherent object discovery in videos.

3 Method

3.1 Preliminary on DINO Attention Maps

The self-supervised ViT, as achieved by DINO [7], has exhibited a rich emergence of objectness in the attention maps. To elucidate this, we visualize the self-attention map from the final Transformer block of the DINO pretrained ViT-S/8 as a representative. As illustrated in Fig. 1(b), an image patch exhibits a high degree of attention dependencies with patches of the same object. Conversely, image patches of distinct objects tend to present different attention distributions. Encouraged by this observation, we pose the following question: Could it be possible to directly harness these attention maps to achieve object segmentation without relying on annotations?

Clustering Metric	Cosine similarity	KL divergence		
	F	A	A_v	$\tilde{A}_v$
DAVIS-16	47.4	60.3	51.1	75.4
DAVIS-17	14.7	26.7	20.9	39.2

Table 1: Preliminary video object segmentation results using varying clustering metrics. F represents features pretrained with DINO. A denotes per-frame attention maps from DINO, while A_v and $\tilde{A}_v$ denote spatio-temporal attention maps produced by a randomly initialized and a final learned temporal correlator respectively, on top of DINO. The IoU ($\mathcal{J}$ score) is reported for both DAVIS-16 and DAVIS-17-Unsupervised datasets.
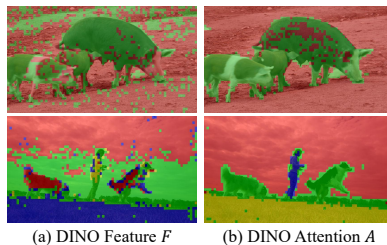

(a) DINO Feature F (b) DINO Attention A

Fig. 2: Visualizations of the clustering results. The left column is the results of DINO features F, the right column is the results of DINO attention A. F results in much noisier clusters, while A distinguishes different classes of objects.

In response to this question, we conduct a straightforward initial exploration with these attention maps. We denote the attention map of the last Transformer block as $A \in \mathbb{R}^{HW \times HW}$ and its output features $F \in \mathbb{R}^{HW \times C}$, where H, W and C respectively denote the height, width, and channel dimension of the feature
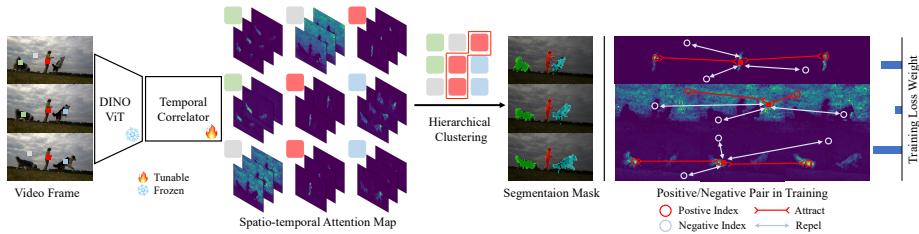
Fig. 3: Our architecture BA overview. Given the video frames and a DINO pretrained Transformer, we first use a temporal correlator to construct the spatio-temporal correspondence. We then utilize these attention maps as a clustering metric and apply hierarchical clustering across all frames to generate segmentation masks. During training, for each patch, we sample a positive/negative set and assign an importance weight based on its corresponding attention map. We promote alignment within the positive set while differentiating the representations from the negative set. The final loss, normalized with importance weights, only trains the temporal correlator with the DINO ViT remaining frozen.

map[5]. For an intuitive comparison, we apply unsupervised clustering on both attention maps A and features F to yield the object segmentation masks. For the detailed procedure, please refer to the Sec. 3.2. Then, we report frame-wise evaluation results on the video object segmentation datasets DAVIS-16 [45] and DAVIS-17-Unsupervised [46] in Table 1. Interestingly, we observe that the attention maps significantly excel in decoupling object components compared to features, exhibiting an improvement of over **10** points on both datasets in terms of IoU. Surprisingly, a spatio-temporal attention map A_v generated by a randomly initialized Transformer block on top of DINO also outperforms the frame-wise DINO feature F. And it is further demonstrated in Fig. 2. The DINO features lead to very noisy clusters, while attention maps can discriminate objects of different categories. We speculate this is because the attention maps preserve intact correspondences, which are more suitable for fine-grained tasks such as object grouping and discrimination. Therefore, creating spatio-temporal structural dependencies that expand beyond spatial DINO attention maps appears promising for video object segmentation tasks.

3.2   Our Simple Architecture for VOS

Motivated by the frame-wise preliminary result, we introduce a temporal correlator instantiated with a spatio-temporal Transformer block [3] on top of the DINO pretrained ViT backbone, to extend the self-attention calculation to the spatio-temporal domain. Specifically, given the frame-wise DINO features $\{F_1, \ldots, F_T\}$ of T frames, we process them through a spatio-temporal Transformer block Φ_{st}. It consists of one standard ViT encoder layer, with self-attention computed over

---

[5] Here we omit the `CLS` token, and average the multi-head attention along the head dimension.

the spatio-temporal grids. Specifically, we concatenate the frame-wise DINO features along the temporal dimension with the addition of learnable temporal positional embeddings $\texttt{pe}$, i.e., $F = \{F_1 + \texttt{pe}_1, \ldots, F_T + \texttt{pe}_T\} \in \mathbb{R}^{THW \times C}$ and then linearly project F to obtain the queries $Q = f_Q(F)$, keys $K = f_K(F)$, and values $V = f_V(F)$. Then, we can obtain the self-attention map $A_v \in \mathbb{R}^{THW \times THW}$ and temporally fused features $F_v \in \mathbb{R}^{THW \times C}$:

$$A_v = \text{softmax}\left(\frac{QK^\top}{\sqrt{C}}\right), \quad F_v = A_v V = \{\tilde{F}_1, \ldots, \tilde{F}_T\}. \tag{1}$$

In this way, we establish spatio-temporal structural dependencies beyond frame-wise DINO features and have the potential to improve object permanence across the entire time span.

Since $A_v[i] \in \mathbb{R}^{THW}$ is a proximity distribution that describes the structural dependencies between the i-th patch and all other positions, we adopt symmetric KL divergence to measure the distance $\mathcal{M} \in \mathbb{R}^{THW \times THW}$ between any two patches as:

$$\mathcal{M}[i,j] = \mathcal{D}_{KL}(A_v[i]||A[j]) + \mathcal{D}_{KL}(A_v[j]||A_v[i]), \tag{2}$$

where $\mathcal{D}_{KL}$ denotes KL divergence. Based on this distance matrix $\mathcal{M}$, we then apply Hierarchical Clustering [28] to obtain the object segmentation masks. Please consult our appendix for a detailed implementation of the hierarchical clustering algorithm. In this way, we iteratively merge THW patches into N centroids $A_c \in \mathbb{R}^{N \times THW}$. Since each centroid likely represents the spatial distribution of an object, we consider the cluster assignments as potential candidates for the object segmentation mask. Specifically, we calculate the KL divergence distance between the A_v and A_c as $D \in \mathbb{R}^{THW \times N}$, and use $\texttt{argmin}$ operation to produce the cluster assignments Z within the video sequence:

$$Z = \texttt{argmin}(D, \texttt{dim=1}) \in \{1, ..., N\}^{THW}. \tag{3}$$

Then, we employ a bipartite matching mechanism, following the standard evaluation protocol in DAVIS-17 Challenge [46], to match the predicted masks with ground truth for evaluation.

A recent study, DiffSeg [56], used hierarchical clustering to process attention maps from a stable diffusion model, but it operated solely in the image domain. Our work extends the powerful image foundation model to the spatio-temporal domain, demonstrating that accurate spatio-temporal correspondence can be learned without supervision. Additionally, our method does not rely on hierarchical clustering. As shown in Table 4, using K-means clustering also achieves competitive performance, indicating that the learned spatio-temporal correspondence is effective across various clustering strategies.

Besides, there are two noteworthy aspects in our clustering stage. First, in the clustering process, we only set a distance threshold hyper-parameter without specifying the number of cluster centroids and the hierarchical clustering algorithm can adaptively generate different numbers of clusters according to

the complexity of the video. Second, our reference to the entire spatio-temporal distribution enables coherent object segmentation across multiple frames. As a result, the generated segmentation masks effortlessly track corresponding objects over time without extra association procedures.

3.3   Training

Noticing that even a randomly initialized attention map A_v shown in Figure 1(c), exhibits an initial ability to differentiate parts and track objects over time, we leverage these maps as indicators for positive and negative sets sampling. Mathematically, given the attention maps $A_v \in \mathbb{R}^{THW \times THW}$, we first `reshape` A_v into shape $THW \times T \times HW$ then use `top-k` operation to retrieve K_p position indexes in each frame with highest attention scores:

$$I_p = \texttt{top-k}(\texttt{reshape}(A_v), K_p, \texttt{dim=-1}). \tag{4}$$

In this way, we obtain a position index set $I_p \in \mathbb{R}^{THW \times T \times K_p}$, which represents positions potentially associated with the same object as the i-th patch. The `reshape` operation ensures that we consistently extract the same number of patches in each frame. This, in turn, guarantees effective object tracking within consecutive frames and promotes consistent temporal correspondence learning. Similarly, we perform this operation to find K_n positions in each frame with the lowest attention scores $I_n \in \mathbb{R}^{THW \times T \times K_n}$. Each $I_n[i]$ represents positions least likely to belong to the same object as the i-th patch. After obtaining the positive and negative index sets, we apply self-supervised correspondence learning in two aspects: semantic consistency and dynamic motion consistency.

Semantic consistency. Intutively, we take the temporally fused features $F_v = \{\tilde{F}_1, ..., \tilde{F}_T\} \in \mathbb{R}^{THW \times C}$ as the semantic representation. We promote the semantic alignment between the features of the same object and distinguish features from different objects. Given the i-th patch, we first refer to the positive pair indexes $I_p[i]$ in Eq. 4 to gather the features that are likely from the same object. Then we calculate the cosine distance between the query feature and all positive feature vectors, and produce the distance matrix $\mathcal{S}_p[i] \in \mathbb{R}^{TK_p}$. Similarly, we gather distinct features using $I_n[i]$, and form negative distance matrix $\mathcal{S}_n[i]$. We employ a simple margin loss to enforce higher consistency between corresponding object areas:

$$\mathcal{L}_s[i] = \sum_{j=1}^{TK_p} \sum_{k=1}^{TK_n} \max(\mathcal{S}_p[i,j] - \mathcal{S}_n[i,k] + \lambda_1, 0), \tag{5}$$

where λ_1 is the margin hyper-parameter and we set $\lambda_1 = 0.8$ as default.

Dynamic motion consistency. Regarding dynamic motions, rather than calculating optical flow, inter-frame feature correlations, or latent cost volumes as motion representations, we turn to highly accessible attention maps A_v. These maps exhibit spatio-temporal correlations of specific patches, serving as an effective latent representation of temporal dynamics. Therefore, it is feasible to

directly employ these attention maps to bolster motion consistency between corresponding objects. Similar to the semantic consistency learning, given the attention map of i-th patch $A_v[i]$ as a query, we also leverage the positive (negative) pair indexes $I_p[i]$ ($I_n[i]$) to gather the attention maps of the corresponding (distinct) objects. Then we calculate the symmetric KL divergence as in Eq. 2 between the query and all positive (negative) attention maps, obtaining the distance matrix $\mathcal{M}_p[i] \in \mathbb{R}^{TK_p}$ and $\mathcal{M}_n[i] \in \mathbb{R}^{TK_n}$. We utilize the same margin loss for optimization, with $\lambda_2 = 1.0$ as the margin hyper-parameter:

$$\mathcal{L}_{\mathrm{m}}[i] = \sum_{j=1}^{TK_p} \sum_{k=1}^{TK_n} \max(\mathcal{M}_p[i,j] - \mathcal{M}_n[i,k] + \lambda_2, 0). \tag{6}$$

Overall objectives. Besides, we evaluate the importance of each spatio-temporal patch and lay more emphasis on the informative regions. Formally, we employ the entropy of the attention maps as a measure of the information content within each patch:

$$e[i] = \sum_{k=1}^{THW} -A_v[i,k] \log A_v[i,k], \tag{7}$$

with $e[i]$ denoting the entropy of i-th patch. A higher entropy value indicates more ambiguous spatio-temporal correlations, which provide less information. Hence, we perform `softmax` normalization on the reversed entropy to produce the importance weight w for each patch as:

$$w[i] = \frac{\exp(-e[i])}{\sum_{j=1}^{THW} \exp(-e[j])}. \tag{8}$$

Then, the overall learning objectives on semantic and motion consistency can be formulated as the weighted summation over all spatio-temporal patches:

$$\mathcal{L} = \sum_{i=1}^{THW} w[i](\mathcal{L}_{\mathrm{s}}[i] + \mathcal{L}_{\mathrm{m}}[i]). \tag{9}$$

Since we only introduce one additional temporal correlator and an entirely parameter-free clustering strategy, we exclusively tune a single standard spatio-temporal Transformer block with 1.6M parameters and freeze the DINO encoder throughout the training process. This straightforward method not only maintains a minimal number of learnable parameters but also yields remarkable results in video object segmentation.

4 Experiments

4.1 Datasets

For training, we adopt the challenging video dataset YouTube-VIS-19 [69] for video instance segmentation. YouTube-VIS-19 consists of 2,883 high-resolution

YouTube videos and each video usually contains multiple object instances of distinct semantics. During the inference, we evaluate our method on unsupervised multiple object segmentation. We use the synthetic MOVi-E [19] dataset and real-world datasets like DAVIS-17-Unsupervised [46] and YouTube-VIS-19 [69], reporting Foreground Adjusted Rand Index (FG-ARI) and mIoU. For MOVi-E, we integrate its training set into our training due to significant distribution shifts. For the rest, we maintain solely training on the YouTube-VIS-19 set. Note that we do not present any mask annotation to the model during either the training or inference stages.

4.2   Implementation Details

In training, we sample $T = 3$ frames with temporal stride 4 as the input clip. Each frame is augmented with random crop and color jitter and resized to 192×384. We adopt DINO pretrained ViT-S/8 [7] as the frame encoder, which is then followed by a single spatio-temporal Transformer Encoder block with 8 heads as temporal correlator Φ_{st}. We set the number of sampled positive and negative pairs to $K_p = 10$ and $K_n = 50$ in default. For optimization, we adopt AdamW [40] with a learning rate 1×10^{-4}, and train the model for a total of 30k iterations with a batch size of 16. In the inference stage, our model can be flexibly applied to video sequences of arbitrary lengths. For inference, we set the distance threshold in hierarchical clustering to 1.0. We follow the standard evaluation protocol outlined in the DAVIS-17 challenge [46] to match the predicted masks with the ground truth.

Table 2: Quantitative results on multiple object video segmentation. For MOVi-E and YouTube-VIS-19 datasets, we report FG-ARI and mIOU. Besides, we report Region Similarity $\mathcal{J}$ and Contour Accuracy $\mathcal{F}$ on DAVIS-17-Unsupervised. ⋆ The original paper of TimeT [52] includes the IoU of the background category, we rerun the evaluation to exclude background here. † We use DINOv2 ViT-S/14 as our frozen frame encoder.

| | MOVi-E | | YTVIS-19 | | DAVIS-17 | | | |
Model	FG-ARI	mIoU	FG-ARI	mIoU	FG-ARI	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SAVi [29]	42.8	16.0	11.1	12.7	-	-	-	-
STEVE [55]	50.6	26.6	20.0	20.9	-	-	-	-
OCLR [65]	-	-	15.9	32.5	14.7	-	34.6	-
VideoSAUR [74]	73.9	35.6	39.5	29.1	-	-	-	-
SOLV [1]	80.8	-	29.1	45.3	32.2	-	30.2	-
SMTC [48]	-	-	31.4	38.8	33.3	40.5	36.4	44.6
TimeT⋆ [52]	-	-	37.9	40.4	35.5	40.0	35.8	44.2
BA (ours)	83.4	40.2	44.3	50.1	40.1	43.9	39.2	48.6
BA† (ours)	84.4	40.7	44.5	50.1	41.6	43.7	39.4	48.0

4.3   Comparison with State-of-the-art

We evaluate our method on challenging multi-object segmentation, which aims to segment different objects in a temporally coherent manner. Among the compared methods, SAVi [29] and OCLR [65] employ optical flow as supervision to discriminate motions of different objects. STEVE [55], VideoSAUR [74], SOLV [1] and SMTC [48] develop variants of slot attention with temporal constraints to bind to distinct instances. TimeT [52] propagates dense DINO features across time to enhance temporal consistency and applies clustering on learned features to produce object segmentations. In contrast, our method directly uses spatio-temporal attention maps for object segmentation. As shown in Table 2, BA achieves state-of-the-art results on both synthetic (MOVi-E) and realistic video datasets (YTVIS-19 and DAVIS-17), indicating promising generalization ability. Specifically, comparing to the prior arts, we gain an approximately 5-point advantage over both VideoSAUR [74] on YTVIS-19 and SMTC [48] and TimeT [52] on DAVIS-17 in terms of FG-ARI. The superior results under our simple architecture reveal the feasibility of utilizing the spatio-temporal dependencies to mine object cues and derive reliable video object segmentation results. Besides, we also report the results using DINOv2 ViT-S/14 as the frame encoder. Despite the larger patch size, our method still achieves competitive performance. This demonstrates the robustness of our approach to pretraining models.
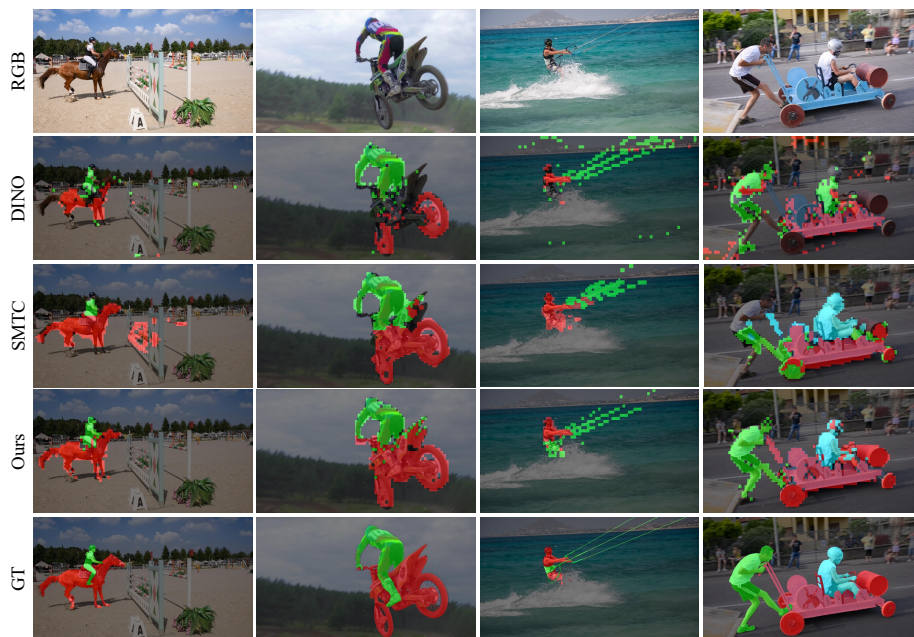

Fig. 4: Qualitative comparison with DINO [7] and SMTC [48] on DAVIS-17-Unsupervised. Different color denotes different clusters.

Qualitative Comparison. To visually demonstrate the effectiveness of our proposed method, we contrast our segmentation results with those of the DINO counterpart and prior arts SMTC [48] in Fig. 4. For the DINO baseline, we cluster the patches through the attention map A. As can be observed, our method is able to segment with significantly less noise. For instance, in the motocross-jump sequence, BA nearly outlines the entire motorbike and the rider, whereas DINO misses most of the bike's body and only identifies its tyre. This shows that inferring the masks by associating a collection of RGB features and constructing spatio-temporal correspondences greatly enhances the segmentation result. Additionally, with the benefit of temporal correspondence, our model can differentiate between distinct object instances with the same semantics. Consider the soapbox sequence displayed in the right column; our method distinguishes two different people - one running behind the soapbox and another sitting in the soapbox throughout the video. This level of distinction cannot be obtained from DINO's spatial attention map, further emphasizing the effectiveness of introducing rich spatio-temporal relations in video segmentation tasks.

4.4   Ablation Study

In the ablation study, we utilize DINO ViT-S/8 as the frozen backbone. We report FG-ARI and IoU on YouTube-VIS-2019 and $\mathcal{J}\&\mathcal{F}$ on DAVIS-17-Unsupervised for multi-object segmentation.

Training objective $\mathcal{L}_\mathbf{s}$, $\mathcal{L}_\mathbf{m}$, and w. To understand the contribution of each component to the VOS performance, we individually evaluate each loss item as depicted in Table 3. Compared to the baseline without training, the inclusion of semantic consistency loss $\mathcal{L}_\mathbf{s}$ with entropy-based weighting mechanism w (second row) elevates FG-ARI from 23.4 to 30.1 on YouTube-VIS-19. Furthermore, the motion consistency loss $\mathcal{L}_\mathbf{m}$ (third row) yields even more substantial improvements, resulting in FG-ARI reaching 39.9. This aligns with our preliminary observations that dynamic motions play a more significant role than semantics in fine-grained object segmentation. Combining all components (fourth row) provides the best results, highlighting the synergy between the components. Additionally, removing the weighting function w (fifth row) slightly decreases the performance on both two benchmarks, affirming the value of entropy normalization to prioritize informative regions.

Clustering metrics and strategy. To further explore whether our clustering strategy optimally leverages the learned spatio-temporal properties, we select four types of clustering metrics for comparison. Specifically, we adopt both feature vectors and attention maps directly from DINO and those from our learned spatio-temporal Transformer block. Table 4 reveals three key observations. First, our learning paradigm consistently outperforms the off-the-shelf DINO in terms of both features and attention maps, improving all metrics by over 10 points. This strongly supports our motivation that establishing spatio-temporal correspondence is crucial for video object segmentation. Second, utilizing the intermediate attention maps directly for clustering proves to be more effective than employing the final features. We speculate this occurs because the attention maps encode

fine-grained structural dependencies between various positions, whereas feature vectors primarily capture high-level semantics. Third, the hierarchical clustering strategy is not mandatory for success. We also achieve similar performance by replacing it with K-means clustering, where we set the number of centroids to 8, indicating that BA distills general object-aware representations across time.

Table 3: Ablation on different combinations of semantic consistency loss $\mathcal{L}_s$, motion consistency loss $\mathcal{L}_m$ and entropy weight w. The first row represents the results obtained from a randomly initialized spatio-temporal Transformer block.

			YTVIS-19		DAVIS-17	
$\mathcal{L}_s$	$\mathcal{L}_m$	w	FG-ARI	mIoU	$\mathcal{J}\&\mathcal{F}$	FG-ARI
-	-	-	23.4	25.5	23.1	18.4
✓	✗	✓	30.1	32.4	29.6	27.5
✗	✓	✓	39.9	46.1	38.6	37.3
✓	✓	✓	44.3	50.1	43.9	40.1
✓	✓	✗	43.1	48.1	42.5	38.1

Table 4: Ablation on different clustering metrics. Features F and attention maps A are directly obtained from DINO. F_v and A_v are generated from our learned spatio-temporal Transformer block. A_v and A_v^* respectively denote the hierarchical clustering and K-means clustering results.

Clustering Metric	YTVIS-19		DAVIS-17
	FG-ARI	mIoU	$\mathcal{J}\&\mathcal{F}$ FG-ARI
F	22.1	23.8	16.1	19.4
F_v	33.9	31.4	33.4	32.1
A	30.1	34.8	29.3	30.5
A_v	44.3	50.1	43.9	40.1
A_v^*	42.5	48.5	41.3	39.3

Number of frames T and temporal stride. To investigate whether a wider temporal receptive field can reinforce temporal coherence and enhance object segmentations, we experiment by varying the number of input frames and temporal stride in training. Our results, displayed in Table 5, reveal that the performance significantly deteriorates when the stride is very small, such as a temporal stride of 1 in the first row. However, it begins to plateau upon reaching 4. This is because a small temporal stride leads to very subtle temporal dynamics in consecutive frames, resulting in a trivial learning task while a larger temporal stride provides richer dynamics, effectively encouraging the model to capture temporal consistency and maintain object permanence. Moreover, adding more frames results in diminishing returns, indicating that a modest temporal receptive field is adequate for sufficient temporal context in video object segmentation.

Motion representations. In training, we use the spatio-temporal attention maps as latent motion representations for dynamic motion alignment $\mathcal{L}_m$. There are some alternatives to formulate this motion representation. Here we explore two variants in Table 6. The one is simple global feature correlation, which directly calculates the dot product between spatio-temporal feature vectors, i.e., $C_g = \tilde{F}\tilde{F}^\top \in \mathbb{R}^{THW \times THW}$. The other is local feature correlation, where we sample a sliding local window with size $T_p \times H_p \times W_p$ for each grid, and obtain the local correlation matrix $C_l \in \mathbb{R}^{THW \times T_p H_p W_p}$. It is obvious that using the self-attention maps as motion representation substantially outperforms the performance of the other two methods. Compared to the global correlation, the

Table 5: Ablation on the number of frames T and temporal stride. When sampling more than 3 frames with a larger stride, the method exhibits a diminishing effect.

Table 6: Ablation on different motion representations in calculating motion consistency loss $\mathcal{L}_m$. We compare using the self-attention matrix, simple local and global feature correlation as the latent motion representation.

		YTVIS-19		DAVIS-17	
T	Stride	FG-ARI	mIoU	$\mathcal{J}\&\mathcal{F}$	FG-ARI
3	1	37.4	43.6	37.1	33.6
3	4	44.3	50.1	43.9	40.1
3	8	44.2	50.2	43.3	40.7
5	4	44.8	50.6	44.3	41.5

	YTVIS-19		DAVIS-17	
MotionRep	FG-ARI	mIoU	$\mathcal{J}\&\mathcal{F}$	FG-ARI
Local Corr.	30.5	33.8	29.6	27.7
Global Corr.	39.1	49.2	36.4	33.9
Self-Attention	44.3	50.1	43.9	40.1

learnable parameters in attention layers enhance the modeling capacity, which enables the self-attention maps to capture more comprehensive spatio-temporal dependencies and serve as a better motion representation. The local correlation only captures temporal dynamics in limited reception fields, restricting the ability for long-term perception.

5 Conclusion

In this paper, we introduce a simple yet powerful approach for self-supervised video object segmentation. Specifically, our proposed architecture introduces only a single spatio-temporal Transformer block that ingests frame-wise DINO features and constructs the spatio-temporal correspondence through the attention layer. During training, we adopt a dual self-supervised consistency objective that encompasses both semantic consistency and dynamic motion consistency, supplemented with entropy normalization. In inference, we utilize hierarchical clustering on spatio-temporal attention maps to generate temporally coherent object segmentation masks. This extremely straightforward approach nevertheless delivers state-of-the-art results on challenging real-world multi-object segmentation tasks such as DAVIS-17-Unsupervised and YouTube-VIS-19.

Limitation. While promising, BA still has room for improvement. Firstly, the current spatio-temporal attention block cannot accommodate very long window-size frames due to the quadratic memory cost. Potentially, adopting memory mechanisms [49, 75] that are capable of handling long-duration data may prove more suitable. Secondly, neither pixel-level constraints nor spatio-temporal dependencies consider the granularity in the video. Incorporating multi-scale features [4] may present a promising direction.

Social Negative Impact. Although our method produces coherent video object segmentation without annotation, it could also enhance deepfake video technologies, making it easier to produce false videos for malicious intent. It is important to regulate the use of such technology and ensure it is leveraged for beneficial purposes.

Acknowledgment

References

1. Aydemir, G., Xie, W., Guney, F.: Self-supervised object-centric learning for videos. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=919tWtJPXe
2. Bao, L., Wu, B., Liu, W.: Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5977–5986 (2018)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
4. Bian, Z., Jabri, A., Efros, A.A., Owens, A.: Learning pixel trajectories with multi-scale contrastive random walks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6508–6519 (2022)
5. Bielski, A., Favaro, P.: Move: Unsupervised movable object segmentation and detection. Advances in Neural Information Processing Systems **35**, 33371–33386 (2022)
6. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 221–230 (2017)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
9. Chen, Y., Liu, Y., Jiang, D., Zhang, X., Dai, W., Xiong, H., Tian, Q.: Sdae: Self-distillated masked autoencoder. In: European conference on computer vision. pp. 108–124. Springer (2022)
10. Choudhury, S., Karazija, L., Laina, I., Vedaldi, A., Rupprecht, C.: Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating Motion. In: British Machine Vision Conference (BMVC) (2022)
11. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H.: Motion-aware contrastive video representation learning via foreground-background merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9716–9726 (2022)
12. Ding, S., Qian, R., Xiong, H.: Dual contrastive learning for spatio-temporal representation. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5649–5658 (2022)

13. Ding, S., Xie, W., Chen, Y., Qian, R., Zhang, X., Xiong, H., Tian, Q.: Motion-inductive self-supervised object discovery in videos. arXiv preprint arXiv:2210.00221 (2022)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), `https://openreview.net/forum?id=YicbFdNTTy`
15. Dutt Jain, S., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3664–3673 (2017)
16. Elsayed, G., Mahendran, A., van Steenkiste, S., Greff, K., Mozer, M.C., Kipf, T.: Savi++: Towards end-to-end object-centric learning from real-world videos. Advances in Neural Information Processing Systems **35**, 28940–28954 (2022)
17. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8554–8564 (2019)
18. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems **22**(3), 1341–1360 (2020)
19. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., Kipf, T., Kundu, A., Lagun, D., Laradji, I., Liu, H.T.D., Meyer, H., Miao, Y., Nowrouzezahrai, D., Oztireli, C., Pot, E., Radwan, N., Rebain, D., Sabour, S., Sajjadi, M.S.M., Sela, M., Sitzmann, V., Stone, A., Sun, D., Vora, S., Wang, Z., Wu, T., Yi, K.M., Zhong, F., Tagliasacchi, A.: Kubric: a scalable dataset generator (2022)
20. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. arXiv preprint arXiv:2311.16933 (2023)
21. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=Fx2SbBgcte`
22. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=SaKO6z6Hl0c`
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
24. Hu, Y., Wang, R., Zhang, K., Gao, Y.: Semantic-aware fine-grained correspondence. In: European Conference on Computer Vision. pp. 97–115. Springer (2022)
25. Hu, Y.T., Huang, J.B., Schwing, A.G.: Videomatch: Matching based video object segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 54–70 (2018)
26. Jabri, A., Owens, A., Efros, A.: Space-time correspondence as a contrastive random walk. Advances in neural information processing systems **33**, 19545–19560 (2020)

27. Johnander, J., Danelljan, M., Brissman, E., Khan, F.S., Felsberg, M.: A generative appearance model for end-to-end video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8953–8962 (2019)

28. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967)

29. Kipf, T., Elsayed, G.F., Mahendran, A., Stone, A., Sabour, S., Heigold, G., Jonschkowski, R., Dosovitskiy, A., Greff, K.: Conditional object-centric learning from video. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=aD7uesX1GF_`

30. Lai, Z., Xie, W.: Self-supervised learning for video correspondence flow. In: BMVC (2019)

31. Lai, Z., Lu, E., Xie, W.: Mast: A memory-augmented self-supervised tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6479–6488 (2020)

32. Li, L., Wang, W., Zhou, T., Li, J., Yang, Y.: Unified mask embedding and correspondence learning for self-supervised video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18706–18716 (2023)

33. Li, X., Loy, C.C.: Video object segmentation with joint re-identification and attention-aware mask propagation. In: Proceedings of the European conference on computer vision (ECCV). pp. 90–105 (2018)

34. Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., Yang, M.H.: Joint-task self-supervised learning for temporal correspondence. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), `https://proceedings.neurips.cc/paper/2019/file/140f6969d5213fd0ece03148e62e461e-Paper.pdf`

35. Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., Yang, M.H.: Joint-task self-supervised learning for temporal correspondence. Advances in Neural Information Processing Systems **32** (2019)

36. Lian, L., Wu, Z., Yu, S.X.: Bootstrapping objectness from videos by relaxed common fate and visual grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14582–14591 (2023)

37. Liu, R., Wu, Z., Yu, S., Lin, S.: The emergence of objectness: Learning zero-shot segmentation from videos. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 13137–13152. Curran Associates, Inc. (2021), `https://proceedings.neurips.cc/paper/2021/file/6d9cb7de5e8ac30bd5e8734bc96a35c1-Paper.pdf`

38. Liu, Y., Wang, Y., Chen, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Promoting semantic connectivity: Dual nearest neighbors contrastive learning for unsupervised domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3510–3519 (June 2023)

39. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in Neural Information Processing Systems **33**, 11525–11538 (2020)

40. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)

41. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and local-

ization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8364–8375 (2022)

42. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

43. Patil, P.W., Dudhane, A., Kulkarni, A., Murala, S., Gonde, A.B., Gupta, S.: An unified recurrent video object segmentation framework for various surveillance environments. IEEE Transactions on Image Processing **30**, 7889–7902 (2021)

44. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2663–2672 (2017)

45. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)

46. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)

47. Qian, R., Ding, S., Liu, X., Lin, D.: Static and dynamic concepts for self-supervised video representation learning. In: European Conference on Computer Vision. pp. 145–164. Springer (2022)

48. Qian, R., Ding, S., Liu, X., Lin, D.: Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16675–16687 (2023)

49. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. arXiv preprint arXiv:2405.16009 (2024)

50. Qian, R., Li, Y., Liu, H., See, J., Ding, S., Liu, X., Li, D., Lin, W.: Enhancing self-supervised video representation learning via multi-level feature optimization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7990–8001 (2021)

51. Rambhatla, S.S., Misra, I., Chellappa, R., Shrivastava, A.: Most: Multiple object localization with self-supervised transformers for object discovery. arXiv preprint arXiv:2304.05387 (2023)

52. Salehi, M., Gavves, E., Snoek, C.G., Asano, Y.M.: Time does tell: Self-supervised time-tuning of dense image representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16536–16547 (2023)

53. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22**(8), 888–905 (2000)

54. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. In: BMVC 2021-32nd British Machine Vision Conference (2021)

55. Singh, G., Wu, Y.F., Ahn, S.: Simple unsupervised object-centric learning for complex and naturalistic videos. Advances in Neural Information Processing Systems **35**, 18181–18196 (2022)

56. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. arXiv preprint arXiv:2308.12469 (2023)

57. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: Proceedings of the European conference on computer vision (ECCV). pp. 391–408 (2018)

58. Wang, N., Song, Y., Ma, C., Zhou, W., Liu, W., Li, H.: Unsupervised deep tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1308–1317 (2019)
59. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)
60. Wang, X., Girdhar, R., Yu, S.X., Misra, I.: Cut and learn for unsupervised object detection and instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3124–3134 (June 2023)
61. Wang, X., Misra, I., Zeng, Z., Girdhar, R., Darrell, T.: Videocutler: Surprisingly simple unsupervised video instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22755–22764 (2024)
62. Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S.X., Crowley, J.L., Vaufreydaz, D.: Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
63. Wang, Y., Li, J., ZHANG, X., Shi, B., Li, C., Dai, W., Xiong, H., Tian, Q.: BarleRIa: An efficient tuning framework for referring image segmentation. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=wHLDHRkmEu`
64. Xie, C., Xiang, Y., Harchaoui, Z., Fox, D.: Object discovery in videos as foreground motion clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9994–10003 (2019)
65. Xie, J., Xie, W., Zisserman, A.: Segmenting moving objects via an object-centric layered representation. In: Advances in Neural Information Processing Systems (2022)
66. Xu, H., Ding, S., Zhang, X., Xiong, H., Tian, Q.: Masked autoencoders are robust data augmentors. arXiv preprint arXiv:2206.04846 (2022)
67. Xu, J., Wang, X.: Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10075–10085 (2021)
68. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7177–7188 (October 2021)
69. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019)
70. Yang, Y., Lai, B., Soatto, S.: Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2826–2836 (June 2021)
71. Yang, Y., Loquercio, A., Scaramuzza, D., Soatto, S.: Unsupervised moving object detection via contextual information separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
72. Ye, V., Li, Z., Tucker, R., Kanazawa, A., Snavely, N.: Deformable sprites for unsupervised video decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2657–2666 (2022)
73. Zadaianchuk, A., Kleindessner, M., Zhu, Y., Locatello, F., Brox, T.: Unsupervised semantic segmentation with self-supervised object-centric representations. In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=1_jFneF07YC`

74. Zadaianchuk, A., Seitzer, M., Martius, G.: Object-centric learning for real-world videos by predicting temporal feature similarities. In: Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023) (2023)

75. Zhou, J., Pang, Z., Wang, Y.X.: Rmem: Restricted memory banks improve video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18602–18611 (2024)

76. Ziegler, A., Asano, Y.M.: Self-supervised learning of object parts for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14502–14511 (2022)