This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Segment Anything Meets Point Tracking

Frano Rajič[1,3] Lei Ke[1,2] Yu-Wing Tai[2] Chi-Keung Tang[2] Martin Danelljan[1] Fisher Yu[1]
[1]ETH Zürich [2]HKUST [3]EPFL

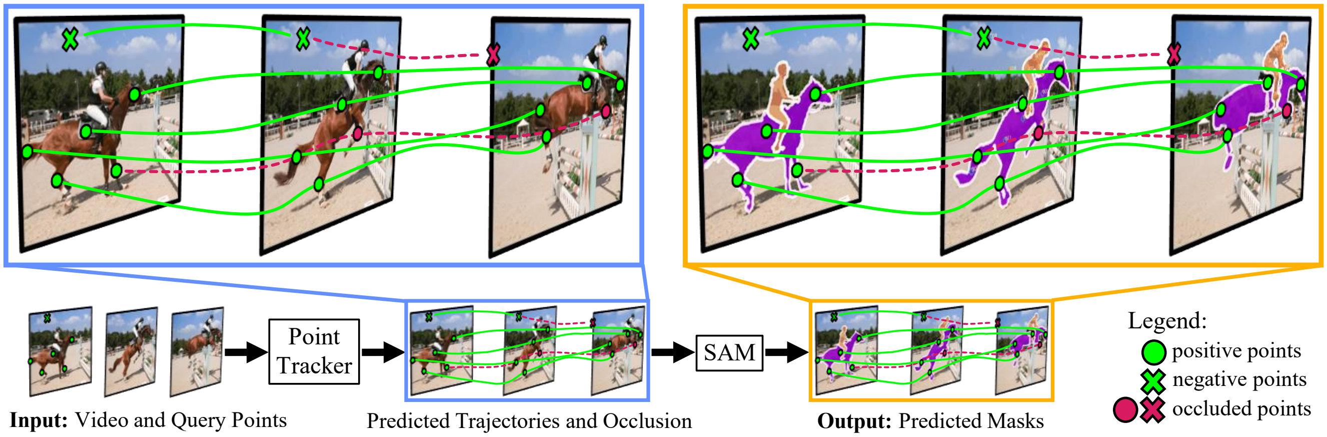

Point Tracker
SAM
Legend:
● positive points
✖ negative points
●✖ occluded points
Input: Video and Query Points
Predicted Trajectories and Occlusion
Output: Predicted Masks

Figure 1. Segment Anything Meets Point Tracking (SAM-PT). SAM-PT is a *point-centric* method that utilizes sparse point propagation for video segmentation. We extend SAM [22] with long-term point trackers to effectively operate on videos in a *zero-shot* manner. SAM-PT takes user clicks as "query points" that denote the target object (positive points) or designate non-target segments (negative points). The points are tracked throughout the video using point trackers that propagate the query points to all video frames, producing trajectory predictions and occlusion scores. SAM is subsequently prompted with the non-occluded points in the trajectories as to output a segmentation mask for each video frame independently. The final masks are refined and optionally used for tracking re-reinitialization, and the propagated points can be further edited for accurate interactive segmentation and tracking.

Abstract

Foundation models have marked a significant stride toward addressing generalization challenges in deep learning. While the Segment Anything Model (SAM) has established a strong foothold in image segmentation, existing video segmentation methods still require extensive mask labeling for fine-tuning, or face performance drops on unseen data domains otherwise. In this paper, we show how foundation models for image segmentation make a step toward enhancing domain generalizability in video segmentation. We discover that, combined with long-term point tracking, image segmentation models yield state-of-the-art results in zero-shot video segmentation across multiple benchmarks. Surprisingly, point trackers exhibit generalization to domains beyond their synthetic pre-training sequences, which we attribute to the trackers' ability to harness the rich local information in the vicinity of each tracked point. Thus, we introduce SAM-PT, an innovative method for point-centric video segmentation, leveraging the capabilities of SAM alongside long-term point tracking. SAM-PT extends SAM's capability to tracking and segmenting anything in dynamic videos. Unlike traditional video segmentation methods that focus on object-centric mask propagation, our approach uniquely exploits point propagation to utilize local structure information independent of object semantics. The effectiveness of point-based tracking is underscored by direct evaluation on the zero-shot open-world UVO benchmark. Our experiments on popular video object segmentation and multi-object segmentation tracking benchmarks, including DAVIS, YouTube-VOS, and BDD100K, suggest that a point-based segmentation tracker yields better zero-shot performance and efficient interactions. We release our code at `https://github.com/SysCV/sam-pt`*.*

1. Introduction

Object segmentation and tracking in videos are central pillars for a myriad of applications, including autonomous driving, robotics, and video editing. Despite significant progress made in the past few years [6,8,45,52], the prevailing methods in semi-supervised Video Object Segmentation (VOS) and Video Instance Segmentation (VIS) exhibit performance gaps when transferred to video domains not seen during training, *i.e.*, in a *zero-shot* setting.

Foundation models have made significant stride in addressing generalization challenges in deep learning. The Segment Anything Model (SAM) [22], trained on 11 million masks and 1 billion object masks, has established itself as the representative foundation model for *image* segmentation, with impressive zero-shot generalization capabilities across tasks in image segmentation. SAM supports being prompted with different modalities, including point prompts, for interactive image segmentation to produce high-quality masks. However, existing methods for *video* segmentation still struggle in zero-shot settings and rely on expensive labels for fine-tuning to achieve high accuracy.

In this work, we show how foundation models for image segmentation make a step toward enhancing domain generalizability in video segmentation. We discover that, combined with *long-term* point tracking, image segmentation models yield state-of-the-art results in zero-shot video segmentation across multiple benchmarks. We have witnessed significant progress in point tracking recently [13–15, 20, 40, 59]. Surprisingly, point trackers exhibit generalization to domains beyond their synthetic pre-training sequences, which we attribute to the trackers' ability to harness the rich local information in the vicinity of each tracked point.

Therefore, we introduce SAM-PT (Segment Anything Meets Point Tracking), the first method to utilize sparse point tracking combined with SAM for video segmentation, offering a new perspective on solving the generalization problem. Instead of employing object-centric dense feature matching or mask propagation, our point-centric approach capitalizes on tracking points using rich local structure information embedded in videos. SAM-PT only requires sparse points annotation to denote the target object in the first frame and provides better generalization to unseen domains. This approach also helps preserve the inherent flexibility of SAM while extending its capabilities effectively to video segmentation. Similar to the integration of SAM in data annotation pipelines, our point-centric approach can potentially be integrated with the existing mask-based approaches in video applications.

SAM-PT prompts SAM with sparse point trajectories as depicted in Fig. 1. These trajectories are predicted by state-of-the-art point trackers, such as CoTracker [20], harnessing their versatility for video segmentation. We found that initializing points to track using K-Medoids cluster centers from a mask label was most compatible with prompting SAM. Tracking both positive and negative points enables clear delineation of target objects from their background. To further refine the output masks, we propose multiple mask decoding passes that integrate both types of points. Additionally, we devised a point reinitialization strategy that increases tracking accuracy over time. This involves discarding points that have become unreliable or occluded, and adding points from object parts or segments that become visible in later frames, such as when the object rotates.

We evaluate SAM-PT on multiple setups including semi-supervised, open-world, and fully interactive video segmentation. Our method achieves stronger performance than existing zero-shot methods by up to 2.5% on DAVIS, 2.0% on YouTube-VOS, and 7.3% on BDD100K, while also surpassing a fully-supervised VIS method [51] on UVO by 5.4 points. We also set up a new benchmark for interactive point-based video segmentation to simulate the process of manually labeling the whole video. In this setup, SAM-PT significantly reduces the annotation effort required for attaining SAM's high-quality segmentation masks for videos and compares favorably to a state-of-the-art interactive method. These results are attained without the need for any video segmentation data during training, underscoring the robustness and adaptability of our approach, and indicating its potential to enhance progress in video segmentation tasks, particularly in zero-shot scenarios.

2. Related Work

Point Tracking for Video Segmentation. Classical feature extraction and tracking methods [2, 26, 27, 36, 38] as well as newer methods [11, 35, 54] have shown proficiency in identifying or tracking sparse features and establishing long-range correspondences. However, these techniques often falter in dynamic, non-rigid environments. While flow-based approaches [37, 47] offer improvements, they too struggle with maintaining long-term point accuracy due to error accumulation and occlusions. Addressing these shortcomings, recent innovations [14, 15, 20, 40, 46, 59] optimize for robust long-term trajectories and effectively manage occlusions. Our work is unique in applying these methods to guide image segmentation models for video segmentation tasks. This is different from methods such as PointTrack [49] that are also point-based but instead use randomly sampled points to encode a global category-specific embedding that can be associated across frames. Ours also differs from other tracking approaches such as by optical flow, box tracking, feature matching, in-context visual prompting, etc., which we elaborate on in Sec. 3.3.

Segment and Track Anything. SAM [22] is a foundation model for image segmentation with impressive zero-shot capabilities. Extensions such as HQ-SAM [21] improve mask quality but are not designed for video tasks. Attempts to extend SAM to video [10, 51] integrate fully-supervised mask trackers [8, 53] but fall short in zero-shot settings.

Zero-Shot VOS and VIS. Generalist models such as Painter [43] apply visual prompting across tasks but show limited VOS performance. SegGPT [44] also uses visual prompting and competes closely with our method on some datasets. Other approaches [5, 18] perform VOS through feature matching. Our approach distinguishes itself by taking the point-centric approach to enhance performance on VOS benchmarks in a zero-shot setting.

Interactive VOS. Interactive VOS has shifted from labor-intensive manual annotations to more user-friendly interaction, such as scribbles, clicks, and drawings, enabling rapid and intuitive video editing [4, 16, 17, 28–30, 39, 55]. MiVOS [9] stands out for its modular design that decouples mask generation from propagation, effectively incorporating user interactions from diverse modalities. Unlike MiVOS and other fully-supervised methods, SAM-PT is the first to use point propagation instead of mask propagation and thus operates effectively in zero-shot settings. Existing benchmarks focus on scribble-based corrections [3] or in-distribution user studies [9], but to fairly compare point-based and brush-based interactions, we set up a new benchmark for interactive point-based video segmentation.

3. Method

We propose SAM-PT for addressing video segmentation tasks in a zero-shot setting. SAM-PT combines the strengths of the Segment Anything Model (SAM), a foundation model for image segmentation, and prominent point trackers, such as PIPS [15] and CoTracker [20], to enable tracking of anything in videos. Sec. 3.1 provides essential background on SAM. Sec. 3.2 details the SAM-PT method with its four constituent steps. Lastly, Sec. 3.3 situates SAM-PT within the current landscape of video segmentation methods as the first point-centric method.

3.1. Preliminaries: SAM

Whereas in computer vision “zero-shot (learning)” usually refers to the study of generalization to unseen object categories in image classification [23], we follow prior work SAM [21, 22, 34] and rather employ the term in a broader sense and explore generalization to unseen datasets.

The Segment Anything Model (SAM) [22] is a novel vision foundation model designed for promptable image segmentation. SAM is trained on the large-scale SA-1B dataset, which contains 11 million images and over 1 billion masks. SA-1B has 400 times more masks than any prior segmentation dataset. This extensive data facilitates SAM’s impressive zero-shot generalization capabilities. SAM has showcased its ability to produce high-quality masks from a single foreground point and has demonstrated robust generalization capacity on a variety of downstream tasks under a zero-shot transfer protocol using prompt engineering. These tasks include, but are not limited to, edge detection, object proposal generation, and instance segmentation.

SAM comprises three main components: an image encoder, a flexible prompt encoder, and a fast mask decoder. The image encoder is a Vision Transformer (ViT) backbone and processes high-resolution 1024×1024 images to generate an image embedding of 64×64 spatial size. The prompt encoder takes sparse prompts as input, including points, boxes, and text, or dense prompts such as masks, and translates these prompts into c-dimensional tokens. The lightweight mask decoder then integrates the image and prompt embeddings to predict segmentation masks in real time, allowing SAM to adapt to diverse prompts with minimal computational overhead.

3.2. Ours: SAM-PT

While SAM shows impressive capabilities in image segmentation, it is inherently limited in handling video segmentation tasks. Our Segment Anything Meets Point Tracking (SAM-PT) approach effectively extends SAM to videos, offering robust video segmentation without requiring training on video segmentation data.

SAM-PT is illustrated in Fig. 2 and is primarily composed of four steps: **1)** selecting query points for the first frame; **2)** propagating these points to all video frames using point trackers; **3)** using SAM to generate per-frame segmentation masks based on the propagated points; **4)** optionally reinitializing the tracking by sampling query points from the predicted masks. We next elaborate on these four steps.

1) Query Points Selection. The process begins with defining query points in the first video frame, which either denote the target object (positive points) or designate the background and non-target objects (negative points). Users can manually and interactively provide query points, or they may be derived from a ground truth mask. For example, in the case of semi-supervised VOS, ground truth mask is provided for the first frame where the object appears. We derive the query points from ground truth masks using different point sampling techniques depicted in Fig. 3 by considering their geometrical locations or feature dissimilarities:

- **Random Sampling:** Randomly selects query points from the ground truth mask.
- **K-Medoids Sampling:** Uses cluster centers from K-Medoids clustering [32] as query points, ensuring good object coverage and robustness to noise.
- **Shi-Tomasi Sampling:** Extracts Shi-Tomasi corner points from the image under the mask, known to be good tracking features [36].
- **Mixed Sampling:** Combines the above techniques to leverage the unique strengths of each.

While each method contributes distinct characteristics influencing performance, K-Medoids performed best in our ablation due to its coverage of various object segments, which helps disambiguate the target object when prompting SAM.

2) Point Tracking. Initiated with the query points, we employ robust point trackers to propagate the points across all frames in the video, resulting in point trajectories and occlusion scores. We adopt point trackers such as PIPS [15] and the state-of-the-art CoTracker [20] to propagate the points as they show moderate robustness toward long-

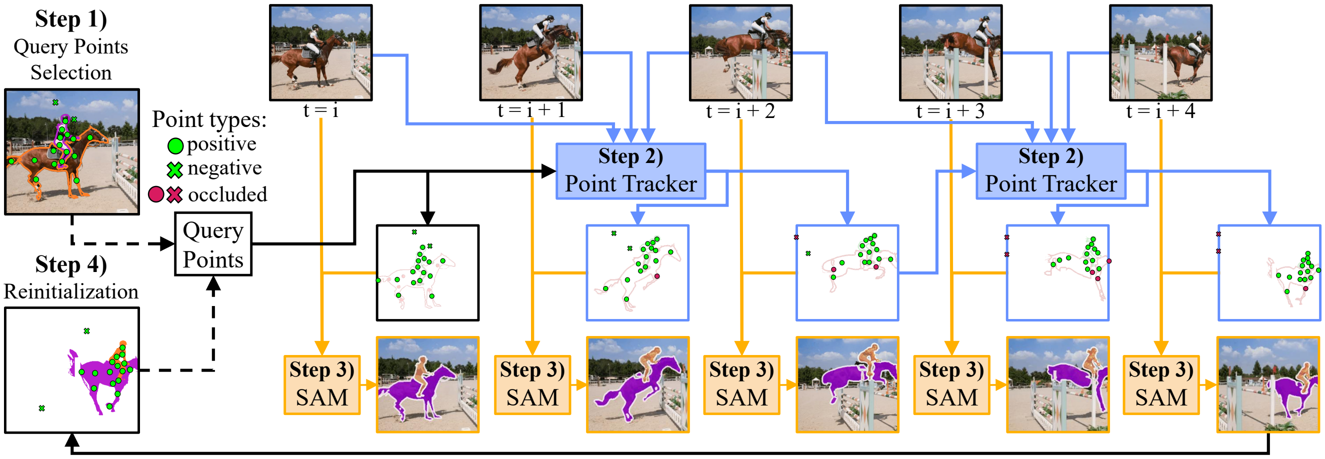


Figure 2. Segment Anything Meets Point Tracking (SAM-PT) overview. The essence of SAM-PT is to extend image segmentation foundation models to effectively operate on videos. SAM-PT has four steps: **1) Query Points Selection.** It starts with first-frame query points which denote the target object (positive points) or designate non-target segments (negative points). These points are provided by the user or derived from a ground truth mask. **2) Point Tracking.** Initiated with the query points, our approach leverages point trackers to propagate the points across video frames, predicting point trajectories and occlusion scores. **3) Segmentation.** The trajectories are then used to prompt the Segment Anything Model (SAM) and output per-frame mask predictions. **4) Point Tracking Reinitialization.** Optionally, the predicted masks are used to reinitialize the query points and restart the process when reaching a prediction horizon h. Reinitialization helps by getting rid of unreliable points and adding points to object segments that become visible in later frames.

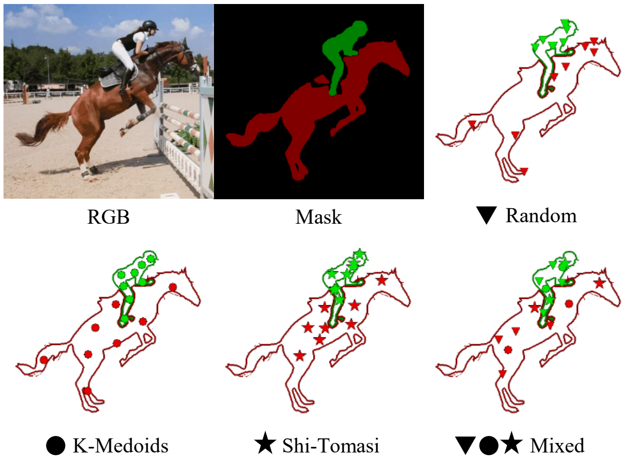


Figure 3. Positive Point Sampling. For an image paired with either a ground truth or predicted segmentation mask, positive points are sampled from within the mask area using one of the following point sampling methods: Random, K-Medoids [32], Shi-Tomasi [36], or Mixed. Notably, Random Sampling and K-Medoids Sampling only require the segmentation mask for input, not the corresponding input image. For negative points, we always use Mixed Sampling on the target object's background mask.

term tracking challenges such as object occlusion and reappearance. Long-term point trackers are also shown more effective than methods such as chained optical flow propagation or first-frame correspondences in our experiments.

3) Segmentation. In the predicted trajectories, the non-occluded points serve as indicators of where the target object is throughout the video. This allows us to use the non-occluded points to prompt SAM, as illustrated in Fig. 4, and leverage its inherent generalization ability to output per-frame segmentation mask predictions. Unlike conventional tracking methods that require training or fine-tuning on video segmentation data, our approach excels in zero-shot video segmentation tasks.

We combine positive and negative points by calling SAM in two passes. In the initial pass, we prompt SAM exclusively with positive points to define the object's initial localization. Subsequently, in the second pass, we prompt SAM with both positive and negative points along with the previous mask prediction. Negative points provide a more nuanced distinction between the object and the background and help by removing wrongly segmented areas.

Lastly, we execute a variable number of mask refinement iterations by repeating the second pass. This utilizes SAM's capacity to refine vague masks into more precise ones. Based on our ablation study, this step notably improves video object segmentation performance.

4) Point Tracking Reinitialization. We optionally execute a reinitialization of the query points using the predicted masks once a prediction horizon of $h = 8$ frames is reached. Upon reaching this horizon, we have h predicted masks and will take the last one to sample new points. At this stage, all previous points are discarded and substituted with the newly sampled points. Following this, steps 1) through 4) are repeated with the new points, starting from the horizon timestep where reinitialization occurs. The steps are iteratively executed until the entire video is processed. The reinitialization process serves to enhance tracking accuracy

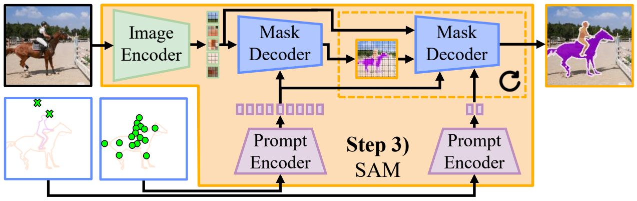


Figure 4. Interacting with SAM in SAM-PT. In the first pass, SAM is prompted exclusively with positive points to define the object's initial localization. In the second pass, both positive and negative points along with the previous mask prediction are fed to the same mask decoder for further mask refinement. The negative points remove segments from the background and neighboring objects and notably help in cases when the point tracker mistakenly predicts positive points off the target object. The second pass is repeated iteratively to get a refined segmentation mask.

over time by discarding unreliable or occluded points while incorporating points from object segments that become visible later in the video. Other reinitialization variants are discussed in our Supplementary Material.

3.3. SAM-PT vs. Object-centric Mask Propagation

By combining sparse point tracking with prompting SAM, SAM-PT distinguishes itself from traditional video segmentation methods that depend on dense object mask propagation, as noted in Tab. 1. To propagate the first-frame GT label to the remaining video frames, traditional techniques commonly use feature matching with masks cached to a mask memory [8, 10, 51, 52], frame-by-frame feature matching [5, 18], optical flow [50], and, recently, in-context visual prompting [43, 44]. In contrast, SAM-PT introduces a unique approach to video object segmentation, employing the robust combination of point tracking with SAM, which is inherently designed to operate on sparse point prompts.

The point propagation strategy of SAM-PT offers several advantages over traditional object-centric tracking methods. First, point propagation exploits local structure context that is agnostic to global object semantics. This enhances our model's capability for zero-shot generalization, an advantage that, coupled with SAM's inherent generalization power, allows for tracking diverse objects in diverse environments, such as on the UVO benchmark. Second, SAM-PT allows for a more compact object representation with sparse points, capturing enough information to characterize the object's segments/parts effectively. Finally, the use of points is naturally compatible with SAM, an image segmentation foundation model trained to operate on sparse point prompts, offering an integrated solution that aligns well with the intrinsic capacities of the underlying model.

SAM-PT stands out as the best-performing method among those not using video segmentation data during training, as outlined in Tab. 1. Although there is a discernible performance gap compared to methods such as XMem [8] that leverage video segmentation training data, our results on zero-shot driving data (Sec. 4.4) and open-world data (Sec. 4.5) show that these methods underperform on unseen data. Furthermore, our method's flexibility extends beyond video object segmentation to tasks such as VIS and interactive point-based video segmentation.

Table 1. Comparative analysis of semi-supervised VOS methods. SAM-PT, introduces *sparse point propagation*, a compact mask representation that uses local structure information agnostic to object semantics. It outperforms other non-video-data-dependent methods, achieving top $\mathcal{J}\&\mathcal{F}$ scores on DAVIS 2016 and 2017, and the highest $\mathcal{G}$ score on YouTube-VOS 2018. The comparison considers the reliance on video mask data during training, the zero-shot learning setting, the initial frame label requirements, and the label propagation techniques used.

Method	**Video Mask**	**Zero-Shot**	**Frame Init.**	**Propagation**	**DAVIS 2016**	**DAVIS 2017**	**YTVOS 2018**
SiamMask [41]	✓	✗	Box	Feature Correlation	69.8	56.4	-
QMRA [25]	✓	✗	Box	Feature Correlation	85.9	71.9	-
TAM [51]	✓	✗	Points	Feature Matching	88.4	-	-
SAM-Track [10]	✓	✗	Points	Feature Matching	92.0	-	-
DEVA [7]	✓	✗	Mask	Feature Matching	-	87.6	-
XMem [8]	✓	✗	Mask	Feature Matching	92.0	**87.7**	86.1
DeAOT [52]	✓	✗	Mask	Feature Matching	**92.9**	86.2	**86.2**
Painter [43]	✗	✓	Mask	Mask Prompting	-	34.6	24.1
STC [18]	✗	✓	Mask	Feature Matching	-	67.6	-
DINO [5]	✗	✓	Mask	Feature Matching	-	71.4	-
SegGPT [44]	✗	✓	Mask	Mask Prompting	82.3	75.6	74.7
HODOR [1]	✗	✓	Mask	Feature Matching	-	77.5	71.7
SAM-PT (ours)	✗	✓	Points	**Points Prompting**	**84.3**	**79.4**	**76.2**

4. Experiments

4.1. Datasets

We evaluate our method on four VOS datasets: DAVIS 2016, DAVIS 2017 [33], YouTube-VOS 2018 [48], and MOSE 2023 [12]. Our interactive point-based video segmentation study also uses DAVIS 2017 and MOSE 2023. We additionally devise a VOS dataset from the BDD100K [56] open driving video dataset. For VIS, we evaluate our method on the class-agnostic dense video instance segmentation task of UVO v1.0 [42]. UVO v1.0 is a VIS dataset aiming for open-world segmentation where objects of arbitrary category are identified and segmented.

4.2. Implementation Details

Training Data. We use pre-trained checkpoints provided by the respective authors for CoTracker [20] and SAM. CoTracker has been trained exclusively on the TAP-Vid-Kubric [13] synthetic data. SAM has been trained on the large-scale SA-1B dataset, the largest image segmentation dataset to date, made up of 11M initially unlabeled images that have been licensed from photographers. HQ-SAM is further trained on the HQ-Seg-44k [21]. Noteworthy, none of these datasets contain video segmentation data, nor do they intersect with any datasets we use for evaluation.

Interactive Point-Based Video Segmentation. To assess the interactive capabilities of SAM-PT, we simulate user refinement of video segmentation results through point interactions. The simulation performs one pass through the video and performs point addition or point removal until a target IoU quality is reached, or a maximum interaction budget is used up. We detail the simulation procedure in the Supplementary Material, including its pseudocode.

4.3. Ablation Study

Our ablation experiments on the DAVIS 2017 validation subset assess different aspects of SAM-PT's design and suggest the following optimal choice of hyperparameters: CoTracker as the point tracker, HQ-SAM variant of SAM, 16 positive and 1 negative point per mask, and 12 refinement iterations. Extended ablation experiments and discussions can be found in the Supplementary Material.

In Tab. 2, we tested SAM-PT with different configurations. We found that using 8 positive points per object instead of just one improved our scores significantly by 33.4 points because one point often was not enough for unambiguously prompting SAM. Selecting points with K-Medoids was slightly better than random and matched Shi-Tomasi, giving a boost of 1.8 points. Incorporating negative points besides positive points helped when trackers made mistakes, such as losing track of an object, and improved scores by another 1.8 points. Adding iterative refinement smoothed out mask quality and fixed some segmentation errors, adding another 2.2 points to the performance. Finally, although reinitialization did not help significantly in the initial tests, it showed benefits on more challenging datasets such as MOSE and UVO, helping in the recovery from tracker errors by discarding incorrect and adding fresh points as well as detecting that the object has disappeared.

The choice of SAM's backbone is important in determining the final performance and inference speed as indicated by Tab. 3. Using the HQ-SAM [21] variant results in the highest performance of 79.4 points, whereas MobileSAM has the highest inference speed of 5.5 FPS. Using lightweight variants doesn't achieve real-time performance as the bottleneck is moved to the point tracker.

4.4. Zero-shot Video Object Segmentation

Performance Overview. SAM-PT sets a new standard in zero-shot VOS on the DAVIS 2017 dataset with a mean $\mathcal{J}\&\mathcal{F}$ score of 79.4, outperforming HODOR's 77.5, SegGPT's 75.6, DINO's 71.4, and Painter's 34.6, as shown in Tab. 4. On the easier DAVIS 2016 validation set, our method achieves 84.3, surpassing SegGPT's 82.3, showcasing the strength of our approach even in less complex scenarios, as detailed in the Supplementary Material.

Table 2. Ablation study on the DAVIS 2017 validation subset of different SAM-PT configurations when using PIPS [15]. We report the mean and std. dev. across eight runs. PSM: point selection method. PP: positive points per mask. NP: negative points per mask. IRI: iterative refinement iterations. R: reinitialization used.

SAM-PT Configuration					**DAVIS [33]**	
PSM	PP	NP	IRI	R	$\mathcal{J}\&\mathcal{F}\uparrow$	Gain
Random	1	0	0	✗	37.1$\pm$21.7	(baseline)
Random	8	0	0	✗	70.5$\pm$1.4	+33.4
Random	16	0	0	✗	70.0$\pm$1.1	
Random	72	0	0	✗	62.6$\pm$0.4	
Shi-Tomasi	8	0	0	✗	72.0$\pm$0.3	
K-Medoids	8	0	0	✗	72.3$\pm$1.2	+1.8
Mixed	8	0	0	✗	70.6$\pm$0.8	
K-Medoids	8	1	0	✗	74.1$\pm$0.7	+1.8
K-Medoids	8	8	0	✗	74.0$\pm$0.8	
K-Medoids	8	16	0	✗	73.4$\pm$0.6	
K-Medoids	8	72	0	✗	72.2$\pm$0.4	
K-Medoids	8	1	1	✗	75.7$\pm$0.7	
K-Medoids	8	1	3	✗	76.0$\pm$0.6	
K-Medoids	8	1	12	✗	76.3$\pm$0.6	+2.2
K-Medoids	8	72	12	✓	**76.8**$\pm$0.7	+0.5

Table 3. Ablation of SAM variants and inference speed when using CoTracker [20] on the DAVIS 2017 validation subset.

SAM Variant	**Backbone**	$\mathcal{J}\&\mathcal{F}\uparrow$	**FPS** $\uparrow$
MobileSAM [58]	ViT-Tiny	71.4$\pm$0.6	**4.6**$\pm$0.2
Light HQ-SAM [21]	ViT-Tiny	72.2$\pm$0.6	4.1$\pm$0.2
SAM [22]	ViT-Base	73.7$\pm$0.7	2.2$\pm$0.1
SAM [22]	ViT-Large	77.5$\pm$0.4	1.6$\pm$0.1
SAM [22]	ViT-Huge	77.6$\pm$0.7	1.2$\pm$0.1
HQ-SAM [21]	ViT-Huge	**79.4**$\pm$0.6	1.1$\pm$0.1

Table 4. DAVIS 2017 validation subset results for semi-supervised VOS. SAM-PT outperforms other zero-shot methods.

Method	**DAVIS 2017 Validation [33]**		
	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
(a) trained on video segmentation data			
AGAME [19]	70.0	67.2	72.7
STM [31]	81.8	79.2	84.3
DeAOT [52]	86.2	83.1	89.2
DEVA [7]	87.6	84.2	91.0
XMem [8]	87.7	84.0	91.4
(b) not trained on any video segmentation data (***zero-shot***)			
Painter [43]	34.6	28.5	40.8
DINO [5]	71.4	67.9	74.9
SegGPT [44]	75.6	72.5	78.6
HODOR [1]	77.5	74.7	80.2
SAM-PT (ours)	**79.4**$\pm$0.6	**76.5**$\pm$0.6	**82.3**$\pm$0.5

However, there is a gap of 8.3 points between SAM-PT and the state-of-the-art fully-supervised XMem, which had been trained on the training data of DAVIS. Despite this gap with in-distribution methods, XMem performs worse compared to SAM-PT when evaluated on zero-shot driving data in Tab. 6 and open-world data in Tab. 7.

In the semi-supervised VOS on YouTube-VOS 2018, we achieve the highest performance among zero-shot methods with 76.2 against SegGPT's 74.7, HODOR's 71.7, and Painter's 24.1, indicating robust generalizability across various video segmentation benchmarks (Tab. 5). On

Table 5. YouTube-VOS 2018 validation subset results for semi-supervised VOS. Metrics are reported separately for "seen" and "unseen" classes, with $\mathcal{G}$ being their overall average score.

Method	YouTube-VOS 2018 Validation [48]				
	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
(a) trained on video segmentation data					
AGAME [19]	66.0	66.9	-	61.2	-
STM [31]	79.4	79.7	84.2	72.8	80.9
RDE [24]	83.3	81.9	86.3	78.0	86.9
XMem [8]	86.1	85.1	89.8	80.3	89.2
DeAOT [52]	86.2	85.6	90.6	80.0	88.4
(b) not trained on video segmentation data (***zero-shot***)					
Painter [43]	24.1	27.6	35.8	14.3	18.7
HODOR [1]	71.7	73.7	76.0	65.5	71.4
SegGPT [44]	74.7	75.1	**80.2**	67.4	75.9
SAM-PT (ours)	**76.2**±0.1	**75.3**±0.1	78.4±0.2	**72.1**±0.2	**79.0**±0.2

Table 6. BDD100K val. subset results for semi-supervised VOS. Metrics include the $\mathcal{J}\&\mathcal{F}$ measure for object visibility durations categorized as short (1-5 frames), medium (6-30 frames), and long (31+ frames). SAM-PT performs better than SegGPT for non-transient objects and than XMem across all except long visibility.

Method	BDD100K VOS Val. [56]					
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$ Short	$\mathcal{J}\&\mathcal{F}$ Med.	$\mathcal{J}\&\mathcal{F}$ Long
(a) trained on video segmentation data but not on BDD100K						
XMem [8]	76.6	74.5	78.7	79.3	78.6	**63.7**
HODOR [1]	78.1	77.5	78.7	90.1	76.1	52.6
(b) not trained on any video segmentation data (***zero-shot***)						
HODOR [1]	67.5	66.9	68.2	78.0	65.4	46.6
SegGPT [44]	**81.5**	**81.2**	**81.8**	**96.1**	78.6	52.0
SAM-PT (ours)	81.0	80.1	**81.8**	91.8	**79.9**	55.8

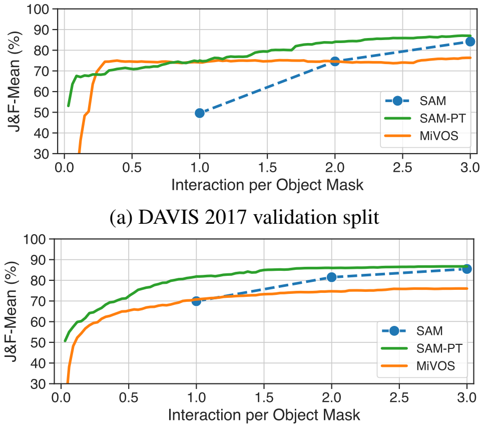
BDD100K's, our method outperforms SegGPT for non-transient objects but also surpasses the fully-supervised XMem across nearly all object visibility durations. The detailed breakdown is provided in Tab. 6. On MOSE 2023, our performance remains competitive with SegGPT, with exact figures available in the Supplementary Material.

Qualitative Analysis. Fig. 6a shows successful segmentation on DAVIS 2017. Our method's ability to perform zero-shot on unseen data is underscored on clips from the anime-influenced series "Avatar: The Last Airbender" in Fig. 7. This highlights its the versatility and adaptability.

Limitations and Challenges. Our method excels in zero-shot VOS but faces challenges with point tracker reliability in complex scenarios, such as occlusions and fast-moving objects, as shown in Fig. 6b. While our point reinitialization and negative point strategies offer improvement, point-based user interactions will quickly recover from point tracking failure cases as suggested by our study in Sec. 4.6.

Table 7. UVO VideoDenseSet v1.0 validation set results. SAM-PT outperforms TAM [51] despite not being trained on any video segmentation data. TAM is a concurrent approach combining SAM and XMem, where XMem was pre-trained on BL30K and trained on DAVIS and YouTube-VOS, but not on UVO.

Method	AR100	ARs	ARm	ARl	AP
(a) trained on video segmentation data, including UVO's training subset					
Mask2Former for VIS [57]	35.4	–	–	–	27.3
ROVIS [57]	41.2	–	–	–	32.7
(b) trained on video segmentation data					
TAM [51]	24.1	21.1	32.9	31.1	1.7
(c) not trained on any video segmentation data (***zero-shot***)					
SAM-PT (ours)	**29.5**	**25.3**	**39.0**	**44.1**	**5.8**



(a) DAVIS 2017 validation split

(b) MOSE 2023 training split

Figure 5. Interactive point-based video segmentation results on **(a)** DAVIS 2017 and **(b)** MOSE 2023, normalized by video length. While SAM and SAM-PT operate in a zero-shot setting, MiVOS [9] was trained on DAVIS 2017. SAM-PT shows more efficient annotation with less user interaction.

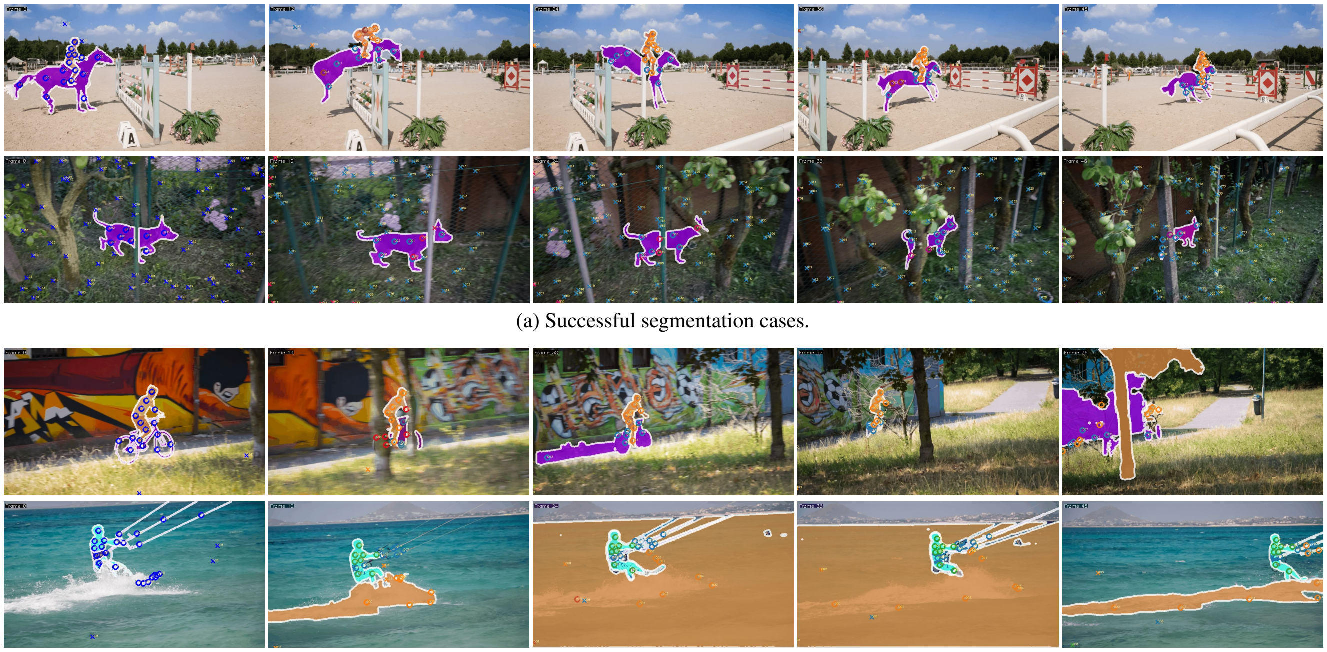
4.5. Open-World Video Instance Segmentation

Tab. 7 suggests that SAM-PT outperforms TAM [51] by 5.4 points given the same mask proposals, despite not being trained on any video segmentation data while TAM was trained on manually annotated DAVIS and YouTube-VOS.

4.6. Interactive Point-Based Video Segmentation

Building upon our SAM-PT's strengths observed in standard benchmarks, this study evaluates the performance of SAM-PT for interactive video annotation tasks. Given that existing brush- and scribble-based benchmarks do not allow for a fair evaluation of point-based interactions, we set up a new benchmark for interactive point-based video segmentation on the DAVIS 2017 [33] and MOSE 2023 [12] datasets.

We benchmarked SAM-PT's interactive performance against a naive SAM approach, which annotates each frame



(a) Successful segmentation cases.

(b) Failure cases with heavy occlusion and thin object structures that lead to point tracking errors.

Figure 6. Visualization of SAM-PT on DAVIS 2017 [33]. Our method segments and tracks objects using the initial object masks from the first frame, with circles denoting positive points and crosses negative points. Red symbols indicate occlusion prediction.

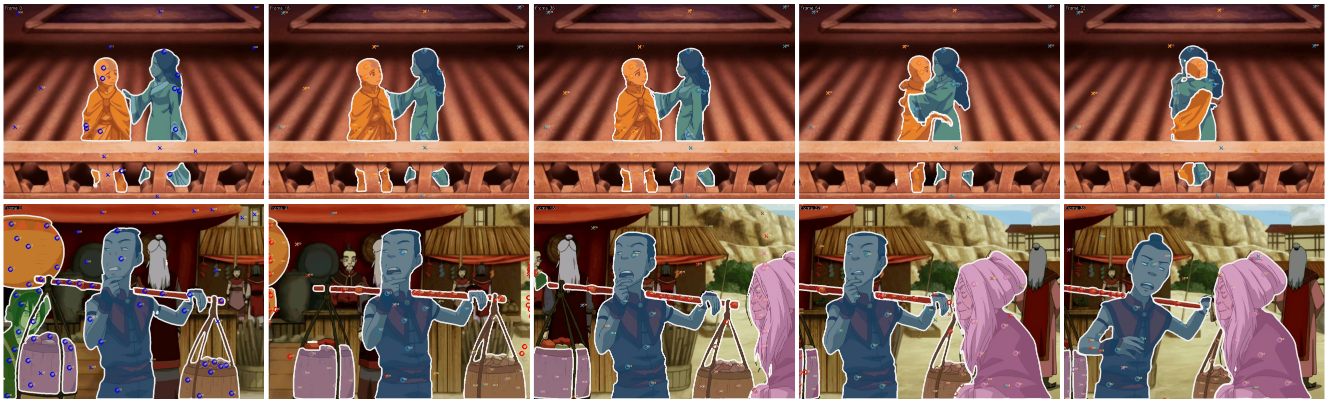


Figure 7. Successful segmentation using SAM-PT on short clips from "Avatar: The Last Airbender". Although our method has never seen data from Avatar, an anime-influenced animated television series, it segments and tracks various objects in short clips.

independently without point tracking, and MiVOS [9], a state-of-the-art method for interactive scribble-based video segmentation. MiVOS supports point interactions but uses mask propagation to propagate interactions to other frames.

Our results are visualized in Fig. 5 and indicate that SAM-PT outperforms baselines, especially on unseen data, reducing the effort required to attain SAM's high-quality masks for videos and highlighting its practical utility.

5. Conclusion

Foundation models have made significant progress toward better generalizability. While SAM shows impressive zero-shot generalization across image segmentation tasks, existing methods for video segmentation still struggle in zero-shot settings and rely on expensive labels for fine-tuning. In this work, we introduce SAM-PT to show how foundation models for image segmentation make a step toward enhancing domain generalizability in video segmentation. Surprisingly, point trackers generalize to domains beyond their synthetic pre-training. SAM-PT achieves strong performance across video segmentation tasks including semi-supervised, open-world, and fully interactive video segmentation. While our method has limitations such as difficulty handling occlusions, small objects, motion blur, and inconsistencies in mask predictions, it contributes a new perspective to video object segmentation research.

References

[1] Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, and Bastian Leibe. Hodor: High-level object descriptors for object re-segmentation in video learned from static images. In *CVPR*, 2022. 5, 6, 7

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 2

[3] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 3

[4] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 5, 6

[6] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1

[7] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 5, 6

[8] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 1, 2, 5, 6, 7

[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, 2021. 3, 7, 8

[10] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 2, 5

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 2

[12] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *CVPR*, 2023. 5, 7

[13] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. In *NeurIPS*, 2022. 2, 5

[14] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. *ICCV*, 2023. 2

[15] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *ECCV*, 2022. 2, 3, 6

[16] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *ECCV*, 2020. 3

[17] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *CVPR*, 2021. 3

[18] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *NeurIPS*, 2020. 2, 5

[19] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *CVPR*, 2019. 6, 7

[20] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *ECCV*, 2023. 2, 3, 5, 6

[21] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 2, 3, 5, 6

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 1, 2, 3, 6

[23] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 3

[24] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *CVPR*, 2022. 7

[25] Fanchao Lin, Hongtao Xie, Yan Li, and Yongdong Zhang. Query-memory re-aggregation for weakly-supervised video object segmentation. In *AAAI*, 2021. 5

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 2

[27] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. 2

[28] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *CVPR*, 2020. 3

[29] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Fast user-guided video object segmentation by interaction-and-propagation networks. In *CVPR*, 2019. 3

[30] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Fast user-guided video object segmentation by interaction-and-propagation networks. In *CVPR*, 2019. 3

[31] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 6, 7

[32] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009. 3, 4

[33] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5, 6, 7, 8

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[35] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[36] Jianbo Shi and Tomasi. Good features to track. In *CVPR*, 1994. 2, 3, 4

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2

[38] Carlo Tomasi and Takeo Kanade. Detection and tracking of point. *IJCV*, 9:137–154, 1991. 2

[39] Jue Wang, Pravin Bhat, R Alex Colburn, Maneesh Agrawala, and Michael F Cohen. Interactive video cutout. In *ACM TOG*, 2005. 3

[40] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *ICCV*, 2023. 2

[41] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 5

[42] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 5

[43] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023. 2, 5, 6, 7

[44] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. In *ICCV*, 2023. 2, 5, 6, 7

[45] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. In *ECCV*, 2022. 1

[46] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *CVPR*, 2024. 2

[47] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, 2022. 2

[48] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 5, 7

[49] Zhenbo Xu, Wei Zhang, Xiao Tan, Wei Yang, Huan Huang, Shilei Wen, Errui Ding, and Liusheng Huang. Segment as points for efficient online multi-object tracking and segmentation. In *ECCV*, 2020. 2

[50] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 5

[51] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 2, 5, 7

[52] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *NeurIPS*, 2022. 1, 5, 6, 7

[53] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *NeurIPS*, 2022. 2

[54] K. M. Yi, Eduard Trulls, Vincent Lepetit, and P. Fua. Lift: Learned invariant feature transform. *ECCV*, 2016. 2

[55] Zhaoyuan Yin, Jia Zheng, Weixin Luo, Shenhan Qian, Hanling Zhang, and Shenghua Gao. Learning to recommend frame for interactive video object segmentation in the wild. In *CVPR*, 2021. 3

[56] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 5, 7

[57] Zitong Zhan, Daniel McKee, and Svetlana Lazebnik. Robust online video instance segmentation with track queries. *arXiv preprint arXiv:2211.09108*, 2022. 7

[58] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 6

[59] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 2