Segment Any Motion in Videos

Nan Huang[1,2] Wenzhao Zheng[1] Chenfeng Xu[1]
Kurt Keutzer[1] Shanghang Zhang[2] Angjoo Kanazawa[1] Qianqian Wang[1]
[1]UC Berkeley [2]Peking University

Abstract

Moving object segmentation is a crucial task for achieving a high-level understanding of visual scenes and has numerous downstream applications. Humans can effortlessly segment moving objects in videos. Previous work has largely relied on optical flow to provide motion cues; however, this approach often results in imperfect predictions due to challenges such as partial motion, complex deformations, motion blur and background distractions. We propose a novel approach for moving object segmentation that combines long-range trajectory motion cues with DINO-based semantic features and leverages SAM2 for pixel-level mask densification through an iterative prompting strategy. Our model employs Spatio-Temporal Trajectory Attention and Motion-Semantic Decoupled Embedding to prioritize motion while integrating semantic support. Extensive testing on diverse datasets demonstrates state-of-the-art performance, excelling in challenging scenarios and fine-grained segmentation of multiple objects. Our code is available at https://motion-seg.github.io/.

1. Introduction

Segmenting moving objects in videos is crucial for a range of applications, including action recognition, autonomous driving [10, 22, 65], and 4D reconstruction [58]. Many prior works address this problem under terms such as Video Object Segmentation (VOS) or motion segmentation. In this paper, we define our task as moving object segmentation (MOS) – segmenting objects that exhibit observable motion within the video. This definition differs from Video Object Segmentation, which includes objects that have the potential to move even if they remain static in the video, and from motion segmentation, which may also capture background motion, such as flowing water. This task is challenging as it implicitly requires distinguishing between camera motion and object motion, robustly tracking objects despite deformations, occlusions, rapid or transient movement, and segmenting them out with precise, clean masks.

Recently, promptable visual segmentation has made significant progress. Taking points, masks, or bounding boxes as prompts, SAM2 [51] segments and tracks the associated objects in videos effectively. However, SAM2 cannot natively handle MOS, as it has no mechanism to detect which objects are moving.

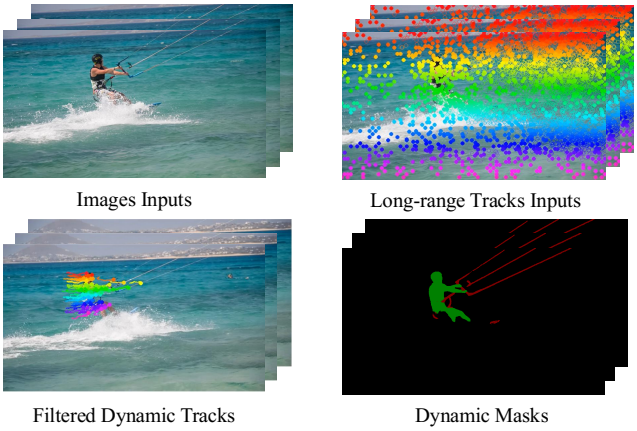

Figure 1. Our method is capable of handling challenging scenarios, including articulated structures, shadow reflections, dynamic background motion, and drastic camera movements, while producing per object level fine-grained moving object masks.

We propose an innovative combination of long-range tracks with SAM2 for moving object segmentation to exploit the capabilities of SAM2. First, point tracking captures valuable long-range pixel motion information which is robust to deformation and occlusion, as shown in Fig. 2. At the same time, we incorporate DINO feature [12, 45], to add semantic context as a complementary source of information to support motion-based segmentation. We depart from traditional MOS approaches by training a model on extensive datasets that effectively combines motion and semantic information at a high level. Given a set of long-range 2D tracks, our model is designed to identify those tracks that correspond to moving objects. Once these dynamic tracks are identified, we apply a sparse-to-dense mask densification strategy, which uses an Iterative Prompting method in conjunction with SAM2 [51] to transform the sparse, point-level mask into a pixel-level segmentation. Since the primary objective is moving object segmentation, we em-

phasize motion cues while using semantic information as secondary support. To effectively balance these two types of information, we propose two specialized modules. (1) Spatio-Temporal Trajectory Attention. Given the long-term nature of input tracks, our model incorporates spatial attention to capture relationships between different trajectories and temporal attention to monitor changes within individual trajectories over time. (2) Motion-Semantic Decoupled Embedding. We implement special attention mechanisms to prioritize motion patterns and process semantic features in supplementary pathways.

We trained our model on extensive datasets, including both synthetic [19, 28] and real-world data [36]. Due to the self-supervised nature of DINO features [45], our model demonstrates strong generalization capabilities, even when primarily trained on synthetic data. We evaluated our approach on benchmarks [34, 43, 47, 48] that were not part of the training data, and the results show that our method significantly outperforms baseline models in diverse tasks.

While previous MOS methods leverage optical flow [6, 8, 9] to capture motion information, either by identifying different motion groups [6, 46, 53, 59] or by using learning-based models [8, 9, 18, 40, 49] to derive pixel masks from optical flow. However, optical flow is limited to short-range motion and can lose track over extended durations. Other methods [3, 7, 14, 42] rely on point trajectories as motion cues, but traditionally utilize spectral clustering on affinity matrices which struggle with complex motions. Though some methods also attempt to take advantage of appearance cues [24, 61] to help understand motion better, they typically handle different modalities in diverse separate stages, limiting the effective integration of their complementary information. Addressing these limitations, our unified framework achieves threefold integration: long-range trajectory, DINO feature, and SAM2. This design explains the model's exceptional capability in handling challenging cases like articulated motion and reflective surfaces as shown in the Fig. 1, and the superior performance in fine-grained segmentation of multiple objects.

In summary, we make the following contributions:

- We introduce an innovative combination of long-range tracks with SAM2, which enables efficient mask densification and tracking across frames.
- To obtain motion labels for trajectories, we propose a method that differs from traditional affinity matrix-based approaches and introduce the Motion-Semantic Decoupled Embedding, which enables a more effective integration of motion and semantic information, enhancing track-level segmentation by balancing these cues.
- Extensive results on multiple benchmarks demonstrate the effectiveness of our method, particularly in fine-grained moving object segmentation.

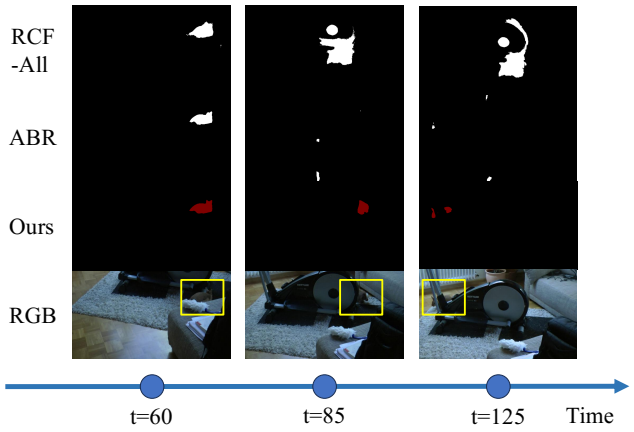

Figure 2. **The effectiveness of long-range tracks**. Over longer periods of time, if a moving object experiences factors such as occlusion or changes in lighting, it can negatively affect the tracking performance of optical-flow-based methods for that object.

2. Related Work

Flow-based Moving Object Segmentation. Traditionally, optical flow based methods [6, 46, 53, 59] segment moving objects by grouping motion cues to create a moving object mask. These methods typically employ iterative optimization or statistical inference techniques to estimate motion models and identify motion regions simultaneously. Recently, numerous deep learning-based approaches [8, 9, 18, 37, 40, 49, 62] have used CNN encoders or transformer to extract motion cues from optical flow, followed by decoders to produce the final segmentation. The main distinctions among these methods lie in model architecture; for instance, methods that encode semantic information often utilize multiple CNN encoders to process different data modalities separately. In general, optical-flow-based methods struggle to distinguish independent object motion from apparent motion caused by depth differences. Furthermore, strong brightness changes also adversely affect these methods. Additionally, optical-flow-based methods are limited to short temporal sequences; they perform poorly if objects move slowly or are occluded.

Trajectory-based Moving Object Segmentation. Trajectory-based methods can be typically classified into two categories: two-frame and multi-frame methods. Two-frame methods [3, 14, 25] generally estimate motion parameters by solving an iterative energy minimization problem, which are recently powered with various convolutional neural network (CNN) models [54, 74]. Multi-frame methods, in contrast, often utilize spectral clustering based on affinity matrices. These matrices are derived through techniques such as geometric model fitting [2, 23, 32, 63], subspace fitting [17, 50, 55, 57], or pairwise motion affinities that integrate motion and appearance information [7, 24, 30, 42]. Recent work has

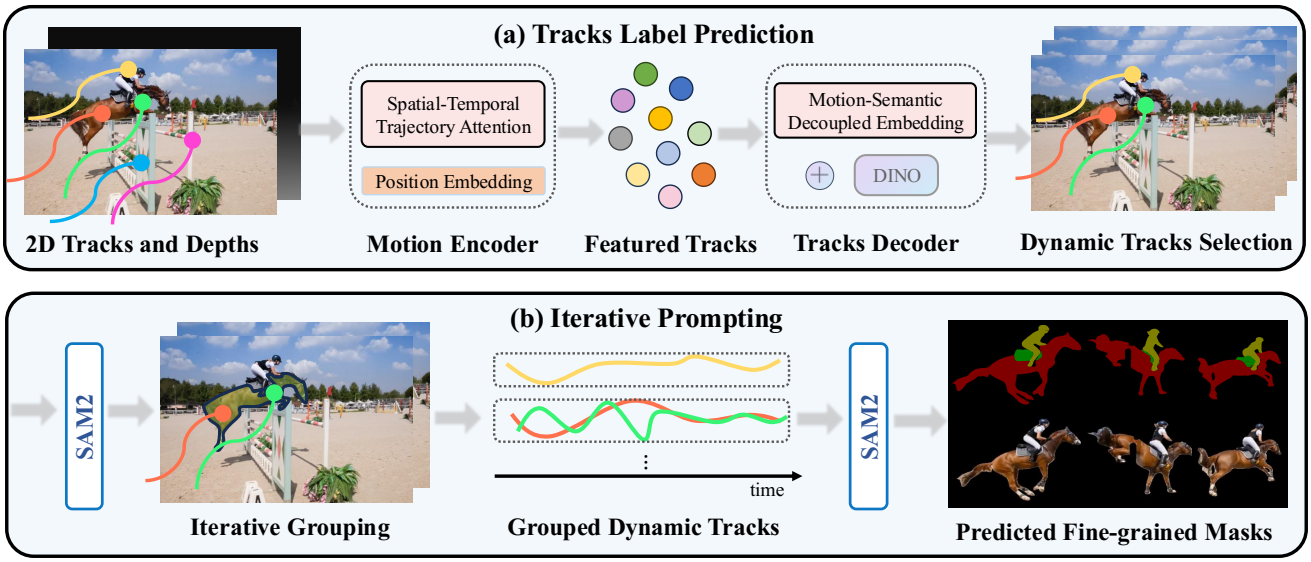

Figure 3. **Overview of Our Pipeline.** We take 2D tracks and depth maps generated by off-the-shelf models [15, 66] as input, which are then processed by a motion encoder to capture motion patterns, producing featured tracks. Next, we use tracks decoder that integrates DINO feature [45] to decode the featured tracks by decoupling motion and semantic information and ultimately obtain the dynamic trajectories(a). Finally, using SAM2 [51], we group dynamic tracks belonging to the same object and generate fine-grained moving object masks(b).

focused on the search for more effective motion models. For instance, [1] uses the trifocal tensor to analyze point trajectories, arguing that it provides more reliable matches over three images than fundamental matrices can over two. However, the trifocal tensor also poses challenges: it is difficult to optimize and prone to failure when the three camera positions are nearly collinear[44]. Other studies [27, 64] have proposed geometric model fusion techniques to combine different models. Some recent work has explored integrating multiple motion cues [21, 41]. For example, [24] investigates combining point trajectories and optical flow, using well-crafted geometric motion models to fuse the two affinity matrices through co-regularized multi-view spectral clustering. However, these approaches still face inherent issues due to their reliance on affinity matrices. They tend to capture only local similarities, leading to poor global consistency, resulting in inconsistent segmentation. Furthermore, affinity matrices is difficult to capture dynamic changes in motion features like speed and direction over time. In contrast, we address the challenge of capturing motion similarities across diverse motion types.

Unsupervised Video Object Segmentation. Unsupervised Video Object Segmentation (VOS) aims to automatically identify and track salient objects in raw video footage, while semi-supervised VOS relies on first-frame ground truth annotations to segment objects in subsequent frames [47, 48]. In this work, we focus on Unsupervised VOS, referred to here simply as "VOS". Recently, many approaches [68, 71] have combined motion and appearance information. For instance, MATNet [73] introduces a motion-attentive transition model for unsupervised VOS, leveraging motion cues to guide segmentation with a primary focus on appearance. RTNet [52] presents a method based on reciprocal transformations, using the consistency of object appearance and motion between consecutive frames to achieve segmentation. FSNet [26] employs a full-duplex strategy with a dual-path network to jointly model both appearance and motion. Overall, VOS generally targets salient objects in videos, regardless of whether the object is moving. Although many VOS methods incorporate motion information, it is often not their primary focus.

3. Method

Our objective is, given a video, to identify moving objects and generate pixel-level dynamic masks. Fig. 3 provides an overview of our pipeline. The central insight is that long-range tracks not only capture motion patterns that facilitate video understanding but also offer long-range prompts essential for promptable visual segmentation. Thus, we use long-range point tracks as motion cues, serving as the primary input in Sec. 3.1, where we apply spatial-temporal attention to capture context-aware feature. In Sec. 3.2, we further incorporate and decouple the use of semantic information with motion cues to decode features, helping the model predict the final motion labels. After identifying dynamic tracks, we leverage these long-range tracks to prompt SAM2 [51] iteratively, as described in Sec. 3.3.

3.1. Motion Pattern Encoding

Point trajectories carry valuable information for understanding motion, and related MOS methods can be typi-

cally classified into two categories: two-frame and multi-frame methods. However, as discussed in Sec. 2, two-frame methods [3, 14, 25] often suffer from significant temporal inconsistencies and exhibit degraded performance when input flows are noisy. Multi-frame methods, in contrast, often utilize spectral clustering based on affinity matrices. Nevertheless, they remain highly sensitive to noise and struggle to handle global, dynamic, and complex motion patterns effectively.

To address these limitations, and inspired by ParticleSFM [72], we propose a method that leverages long-range point tracks [15], processed through a specialized trajectory processing model, to predict per-trajectory motion labels. As illustrated in Fig. 3, our proposed network adopts an encoder-decoder architecture. The encoder directly processes long-range trajectory data and applies a Spatio-Temporal Trajectory Attention mechanism across trajectories. This mechanism integrates both spatial and temporal cues, capturing both local and global information across time and space, in order to embed the motion pattern of each trajectory.

Given that the accuracy and quality of long-range trajectories significantly impact model performance, we utilize BootsTAP [15] to generate the tracks, which provides a confidence score for each track at each time step, enabling us to mask out low-confidence points. Furthermore, due to the movement of dynamic objects and camera motion, the visibility of long-range tracks can vary over time, as they may be occluded or move out of the frame. This variability in visibility and confidence makes each trajectory data highly irregular, motivating our use of a transformer model, inspired by sequence modeling approaches in natural language processing [56, 72], to handle the data effectively.

Our input data comprises long-range trajectories, with each trajectory consisting of normalized pixel coordinates (u_i, v_i), visibility ρ_i and confidence scores c_i, where $i \in (0, \text{time})$. Masks $\mathcal{M}_i$ is applied to indicate points where the pixel coordinates are either invisible or low-confidence. Additionally, we integrate monocular depth maps d_i estimated by Depth-Anything [66], which, despite some noise, provide valuable insights into the underlying 3D scene structure, enhancing understanding of spatial layout and occlusions. To further enrich the input data and strengthen temporal motion cues, we compute frame-to-frame differences in both trajectory coordinates $(\Delta u_i, \Delta v_i)$ and depth Δd_i for adjacent frames.

Since adjacent sampling points in coordinates can lead to oversmoothing of spatially close features, we draw inspiration from NeRF [39] to address this issue. Specifically, we apply frequency transformations for positional encoding to better capture fine-grained spatial details.

The final augmented trajectories pass through two MLPs to generate intermediate features, which are then fed into the transformer encoder. Given the long-range nature of the input data, we propose a Spatio-Temporal Trajectory Attention for our encoder $\mathcal{E}$, interleaves attention layers that operate alternately across track and temporal dimensions [4, 29]. This design allows the model to capture both the temporal dynamics within each trajectory and the spatial relationships across different trajectories. Finally, to obtain a feature representation for each entire trajectory rather than individual points, we perform max-pooling along the temporal dimension, following [72]. This process yields a single feature vector for each trajectory, naturally forming a high-dimensional featured track that implicitly captures the unique motion pattern of each trajectory.

3.2. Per-trajectory Motion Prediction

Though we encoded motion pattern in Sec. 3.1, it is still challenging to distinguish moving objects based solely on motion cues, because learning to differentiate between object motion and camera motion from highly abstracted trajectories is difficult for the model. Providing the model with texture, appearance, and semantic information can simplify this task by helping it understand which objects are likely to move or be moved. Some approaches directly apply semantic segmentation models [5, 20, 67, 70] where potentially moving pixels are identified based on semantic labels. While these methods can be effective in specific scenarios, they are intrinsically limited for general moving object segmentation, as they depend entirely on predefined semantic classes. Recently, many MOS [61, 69] and VOS [11, 33, 35] methods combine appearance information and motion cues, but they do so in two separate stages, often using RGB images to refine masks. However, relying on raw RGB data may fail to capture high-level information, and applying the two modalities in separate stages limits the effective integration of their complementary information.

To address these limitations, we incorporate DINO features predicted by DINO v2 [45], a self-supervised model, which helps generalize the inclusion of appearance information. However, we observed that simply introducing DINO features as input makes the model overly reliant on semantics as shown in Fig. 8 and discussed in Sec. 4.5, reducing its ability to differentiate between moving and static objects within the same semantic category. To overcome this issue, we propose a Motion-Semantic Decoupled Embedding, enabling the transformer decoder $\mathcal{D}$ to prioritize motion information while still considering semantic cues.

We obtain the final embedded featured tracks $\mathcal{P}$ through the process described in Sec. 3.1:

$$\mathcal{P} = \mathcal{E}((\gamma(u), \gamma(v), \gamma(\Delta u), \gamma(\Delta v), d, \Delta d, \rho, c), \mathcal{M}). \quad (1)$$

We then design a transformer-based decoder, where the encoder layer performs attention only on the embedded featured tracks, which contain motion information exclusively.

After computing the attention-weighted feature, we concatenate the DINO feature and pass this concatenated feature through a feed-forward layer. In the decoder layer, self-attention is still applied only to the motion features; however, multi-head attention is used to attend to a memory that includes semantic information. Finally, we apply a sigmoid activation function to produce the final output, yielding the predicted label for each trajectory.

We then compute the loss between these predicted labels and per-track ground truth labels using a weighted binary cross-entropy loss [72]. We assign ground truth labels to each trajectory by checking if the sampled point coordinates lie within the ground truth dynamic masks. If a point falls inside the mask, it is labeled as dynamic.

3.3. SAM2 Iterative Prompting

As depicted in Fig. 3, after obtaining the predicted label of each trajectory and filter dynamic trajectories, we use these trajectories as point prompts for SAM2[51] with an iterative, two-stage prompting strategy. The first stage focuses on grouping trajectories belonging to the same object and storing the trajectories of each distinct object in memory. In the second stage, this memory is used as a prompt for SAM2 [51] to generate dynamic masks.

The motivation behind this approach is twofold. First, it is necessary because SAM2 requires object IDs as input. However, if we assign the same object ID to all dynamic objects (e.g., assigning 1 to represent all dynamic objects), SAM2 would struggle to simultaneously segment multiple objects that share the same ID. Second, this method offers the benefit of achieving finer-grained segmentation.

In the first stage, we select the time frame with the maximum number of visible points and locate the densest point among all visible points in that frame. This point serves as the initial prompt for SAM2 [51], which then generates an initial mask for that frame. After generating this mask, we apply dilation to expand its boundaries, excluding all points within the expanded mask area to remove edge points and assume that these points belong to the same object. We then proceed to the next frame with the highest number of visible points and repeat this process until the remaining visible points across all frames are too few to process. The trajectories identified as belonging to the same object are stored in memory, with unique object IDs assigned to each. We only save the points within the undilated mask for each object.

In the second stage, we use this memory to refine prompt selection by locating the densest point within the stored trajectories and the two points furthest from this point. Leveraging the long-range nature of trajectories, we prompt SAM2 at regular intervals to prevent it from losing track of the object over extended distances. Since SAM2 may generate partial object masks (e.g., parts of a person's clothing), we perform post-processing on all masks to merge those that overlap internally or appear within the same mask boundaries. This results in a complete mask for each distinct object.

4. Experiments

4.1. Implementation Details

Training Dataset. We train our model using three datasets: Kubric[19], Dynamic Replica[28], and HOI4D [36], sampling them at a ratio of 35%, 35%, and 30% respectively. Kubric [19] is a synthetic dataset composed of sequences of 24 frames showing 3D rigid objects falling under gravity and bouncing. We generate dynamic masks for each sequence based on the motion labels of individual objects. Dynamic Replica [28] is another synthetic dataset, created for 3D reconstruction, that includes long-term tracking annotations and object masks, featuring articulated models of humans and animals. We calculate dynamic masks by analyzing the 3D tracks to determine whether each object is in motion, providing accurate motion segmentation for this dataset. HOI4D [36] is a real-world, egocentric dataset that contains common objects involved in human-object interactions. This dataset provides official motion segmentation masks, making it ideal for real-world training of our model.
Data Sampling. During training, we randomly sample a variable number of tracking points, enhancing the model's robustness to different track counts. For the Dynamic Replica dataset [28], which contains 300 frames, we speed up training by sampling 1/4 of the frames at regular intervals randomly. This approach preserves the large camera motion characteristics of the dataset. We find that including the Dynamic Replica dataset is essential for helping the model understand camera motion effectively.

4.2. Benchmark and metrics

We evaluate our model using several established datasets for moving object video segmentation. DAVIS17-Moving[13] is a subset of the DAVIS2017 dataset[48], designed specifically for moving object detection and segmentation. In DAVIS17-Moving, all moving instances within each video sequence are labeled, while static objects are excluded. Following the same criteria, we created DAVIS16-Moving as a subset of the DAVIS2016 dataset [47]. Additionally, we report performance on other popular video object segmentation benchmarks, including DAVIS2016[47], SegTrackv2[34], and FBMS-59 [43].

For evaluation, we benchmark our moving object video segmentation performance using region similarity (J) and contour similarity (F) metrics, as outlined in [35, 38, 61].

4.3. Moving Object Segmentation

We selected methods that specifically target moving object segmentation as baselines [35, 38, 60, 61, 69]. For

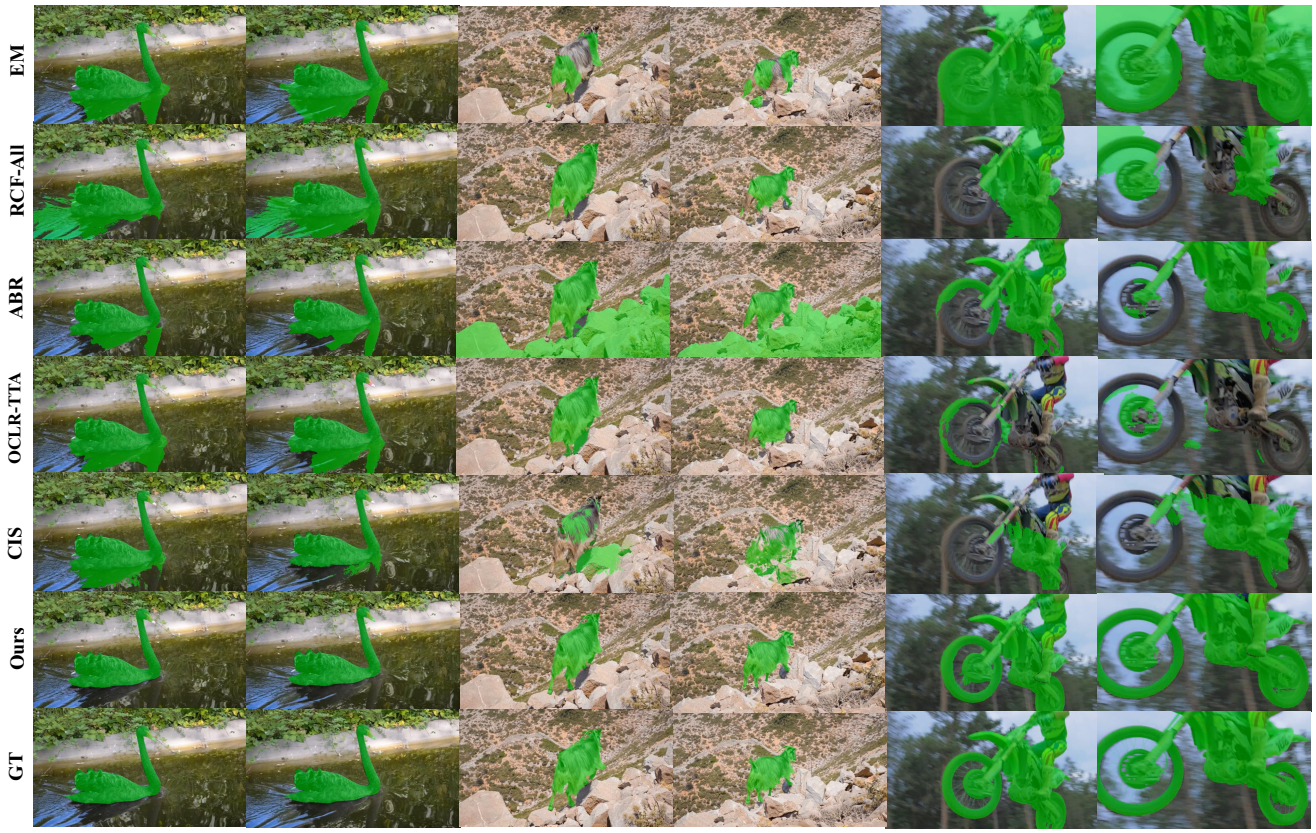

Figure 4. Qualitative comparison on DAVIS17-moving benchmarks. For each sequence we show moving object mask results. Our method successfully handles water reflections (left), camouflage appearances (middle), and drastic camera motion (right).

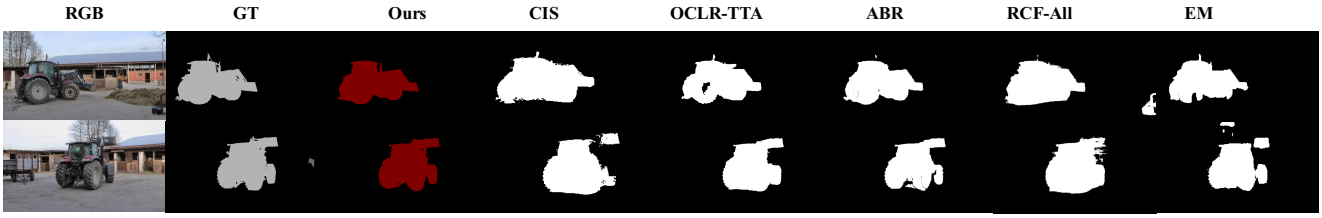

Figure 5. Qualitative comparison on FBMS-59 benchmarks. The masks produced by us are geometrically more complete and detailed.

OCLR [60], we report results for two versions: OCLR-flow, which uses only flow input, and a second version OCLR-TTA that incorporates test-time adaptation on top of OCLR-flow. For RCF [35], the first stage, RCF-stage 1, focuses on motion information, while the second stage, RCF-All, further optimizes the results from the first stage. We report results for both stages. For all baselines, we apply a fully connected conditional random field (CRF) [31] to refine the masks and achieve the best possible results.

Notably, for multi-object scenarios, we follow the common practice [16, 60, 61, 69] of grouping all foreground objects together for evaluation purposes, which we refer to as MOS. Although our approach is capable of generating highly accurate, fine-grained per-object masks, as detailed in Sec. 4.4, we term this second evaluation method as fine-grained MOS. Table 1 compares the performance of our model with several baseline methods on the MOS task. Our method achieves state-of-the-art F-scores across all datasets, and our region similarity (J) scores are either the best or second-best across multiple datasets, further validating the effectiveness of our approach. Fig. 4 shows our visual results on the DAVIS16-Moving dataset, where our method accurately identifies object boundaries without incorrectly labeling moving backgrounds. Moreover, our masks exhibit strong geometric structure, particularly in challenging scenarios with significant camera motion. Fig. 5 and Fig. 6 present qualitative results on the FBMS-59 and SegTrack v2 benchmarks, respectively. Our method performs exceptionally well in maintaining mask geometry, and even in cases where the RGB images are blurred or of

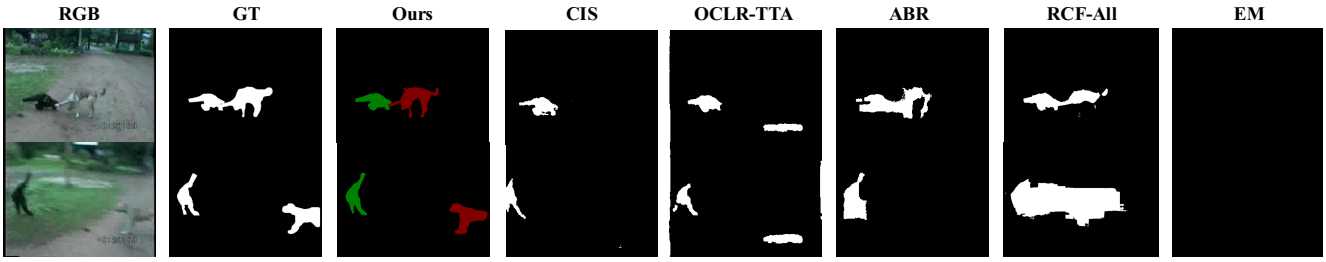

Figure 6. Qualitative comparison on SegTrack v2 benchmarks. Our method succeeds even under motion blur conditions.

Table 1. Quantitative comparison on MOS task which grouping all foreground objects together for evaluation.

Methods	Model Settings		DAVIS2016-Moving			SegTrackv2	FBMS-59		DAVIS2016		
	Motion	Appearance	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
CIS [69]	Optical Flow	RGB	66.2	67.6	64.8	62.0	63.6	-	68.6	70.3	66.8
EM [38]	Optical Flow	✗	75.2	76.2	74.3	55.5	57.9	56.0	70.0	69.3	70.7
RCF-Stage1 [35]	Optical Flow	✗	77.3	78.6	76.0	76.7	69.9	-	78.5	80.2	76.9
RCF-All [35]	Optical Flow	DINO	79.6	81.0	78.3	**79.6**	72.4	-	80.7	82.1	79.2
OCLR-flow [60]	Optical Flow	✗	70.0	70.0	70.0	67.6	65.5	64.9	71.2	72.0	70.4
OCLR-TTA [60]	Optical Flow	RGB	78.5	80.2	76.9	72.3	69.9	68.3	78.8	80.8	76.8
ABR [61]	Optical Flow	DINO	72.0	70.2	73.7	76.6	**81.9**	79.6	72.5	71.8	73.2
Ours	Trajectory	DINO	**89.5**	**89.2**	**89.7**	76.3	78.3	**82.8**	**90.9**	**90.6**	**91.0**

Figure 7. Qualitative comparison on Fine-grained MOS task which will produce per-object level masks.

low quality, our reliance on long-range trajectories enables accurate identification of moving objects.

4.4. Fine-grained Moving Object Segmentation

Building on the initial MOS task, this task not only identifies moving objects but also classifies them within their motion context to generate fine-grained, per-object masks. We evaluate our approach for multi-moving object segmentation specifically on the DAVIS2017-Moving dataset. For a fair comparison, we only include baselines that claim the ability to perform this task. Table 2 shows that our method significantly outperforms the baselines, demonstrating its superior capability in producing accurate per-object masks. Additionally, Fig. 7 illustrates that, first, our method accurately identifies each object, effectively distinguishing different objects with similar motion patterns. Second, it ensures the completeness of each object mask, handling challenging cases such as articulated human structures and occluded objects while maintaining mask integrity.

Table 2. Quantitative comparison on DAVIS17-Moving dataset for MOS and Fine-grained MOS tasks.

Methods	MOS		Fine-grained MOS		
	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
OCLR-flow [60]	69.9	70.0	44.4	42.1	46.8
OCLR-TTA [60]	76.0	75.3	49.1	48.4	49.9
ABR [61]	74.6	75.2	51.1	50.9	51.2
Ours	**90.0**	**89.0**	**80.5**	**77.4**	**83.6**

4.5. Ablation Study

We investigate the effectiveness of our method and its various components on the DAVIS17-Moving and DAVIS16-Moving datasets. The former is used for fine-grained MOS, while the latter focuses on MOS. All models are trained for a full number of epochs.

We conducted several experiments to assess the importance of each component. The w/o DINO configuration ex-

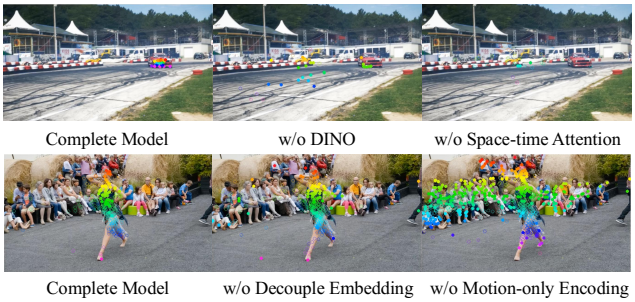

Figure 8. Visual comparison for the ablation study on two critical and challenging cases. The top sequence shows scenarios involves drastic camera motion and complex motion patterns, while the bottom sequence with both static and dynamic objects of the same category. The experimental setup is detailed in Sec. 4.5.

Table 3. Quantitative comparison for the ablation study on the DAVIS17-Moving and DAVIS16-Moving benchmarks, which evaluate fine-grained MOS and MOS tasks, respectively. The experimental setup is detailed in Sec. 4.5.

Methods	DAVIS17-Moving			DAVIS16-Moving		
	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
w/o Depth	69.2	65.6	72.8	82.5	78.6	86.4
w/o Tracks	19.6	14.5	24.7	20.9	9.8	31.9
w/o DINO	65.0	62.1	67.9	75.5	71.4	79.5
w/o MOE	72.0	68.7	75.4	81.8	81.0	82.7
w/o MSDE	63.0	59.3	66.7	78.2	77.3	79.1
w/o PE	66.4	64.7	68.2	82.0	81.5	82.5
w/o ST-ATT	65.5	61.9	69.1	78.3	74.3	82.4
Ours-full	**80.5**	**77.4**	**83.6**	**89.1**	**89.0**	**89.2**

cludes DINO features entirely during training, while w/o MOE (Motion-only Encoding) concatenates DINO features with motion cues before the motion encoder, allowing both the encoder and decoder layers to incorporate DINO information throughout. w/o MSDE (Motion-Semantic Decoupled Embedding) excludes DINO features from the motion encoder but concatenates them with the embedded featured tracks from the encoder output, introducing semantic information through self-attention in the tracks decoder. We also test configurations w/o depth and w/o tracks, removing specific inputs to observe their impact on performance. Additionally, w/o PE (Positional Embedding) omits NeRF-like positional embedding in the motion encoder, and w/o ST-ATT (Spatial-temporal Attention) replaces spatial-temporal attention with conventional attention.

Table 3 presents the quantitative results. We find that excluding depth as input or positional encoding impacts performance less than other components, but it still falls significantly short of the best results. When tracks are removed and only DINO features and depth maps are used, performance drops drastically, indicating that the model struggles to learn effectively without trajectory-based information. We further analyze the key components in two challenging scenarios presented below.

Drastic Camera Motion. We observed that in highly challenging scenes, such as those with drastic camera movement or rapid object motion, relying solely on motion information is insufficient. As shown in the upper part of Fig. 8, the colored points represent dynamic points predicted by the model, while the hollow points indicate invisible points at that moment. In this example, without DINO feature information, the model incorrectly classifies the stationary road surface as dynamic, despite the fact that the road lacks the ability to move. This information can be effectively supplemented by incorporating DINO features. Additionally, we found that adding spatial-temporal attention within the motion encoder is particularly beneficial in these difficult scenarios, as it provides the model with richer motion information to capture the long-range motion patterns of tracks, as illustrated in Fig. 8.

Distinguishing Moving and Static Objects of the Same Category. Results show that excluding DINO features entirely results in a performance drop, and the manner in which these features are integrated significantly affects the model's output. Simply incorporating DINO as an input during the motion encoding stage causes the model to rely heavily on semantic information, often leading it to assume that objects of the same type share the same motion state. In contrast, our Motion-Semantic Decoupled Embedding architecture effectively reduces this over-reliance on semantics, allowing the model to differentiate between moving and static objects within the same category, as illustrated in the lower part of Fig. 8.

5. Conclusion

In this work, we present a novel approach that leverages long-range tracks which departs from traditional affinity matrix-based methods. Trained on extensive datasets, our model accurately identifies dynamic tracks, which, when combined with SAM2, produce precise moving object masks. Our carefully designed model architecture is tailored to handle long-range motion information while effectively balancing motion and appearance cues. Experiments show that our method achieves state-of-the-art results across multiple benchmarks, with particularly strong performance in per-object-level segmentation.

6. Limitation

Our current model utilizes off-the-shelf long-range tracking estimators, whose accuracy can greatly influence overall performance, as shown in Tab. 3. However, we believe that our novel approach to moving object segmentation can, in turn, enhance the performance and understanding of these estimators, creating a mutually beneficial improvement.

References

[1] Federica Arrigoni, Luca Magri, and Tomas Pajdla. *On the Usage of the Trifocal Tensor in Motion Segmentation*, page 514–530. 2020. 3

[2] Federica Arrigoni, Luca Magri, and Tomas Pajdla. *On the Usage of the Trifocal Tensor in Motion Segmentation*, page 514–530. 2020. 2

[3] Daniel Barath and Jiri Matas. Progressive-x: Efficient, anytime, multi-model fitting algorithm. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, 2019. 2, 4

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *Cornell University - arXiv,Cornell University - arXiv*, 2021. 4

[5] Berta Bescos, Jose M. Facil, Javier Civera, and Jose Neira. Dynaslam: Tracking, mapping and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, page 4076–4083, 2018. 4

[6] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. *Cornell University - arXiv,Cornell University - arXiv*, 2016. 2

[7] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer, 2010. 2

[8] M. Bösch. Deep learning for robust motion segmentation with non-static cameras. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, 2021. 2

[9] Zhe Cao, Abhishek Kar, Christian Hane, and Jitendra Malik. Learning independent object motion from unlabelled stereoscopic videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[10] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv:2311.18561*, 2023. 1

[11] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Dogyoon Lee, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Dual prototype attention for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19238–19247, 2024. 4

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 1

[13] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5

[14] Andrew Delong, Anton Osokin, Hossam N. Isack, and Yuri Boykov. Fast approximate energy minimization with label costs. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. 2, 4

[15] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, and Andrew Zisserman. BootsTAP: Bootstrapped training for tracking any point. *arXiv*, 2024. 3, 4

[16] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. 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*, pages 3664–3673, 2017. 6

[17] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 2765–2781, 2013. 2

[18] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. Epo-net: Exploiting geometric constraints on dense trajectories for motion saliency. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, 2019. 2

[19] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 2, 5

[20] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 386–397, 2020. 4

[21] Christian Homeyer and RobertBosch Gmbh. On moving object segmentation from monocular video with transformers. 3

[22] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. S3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 1

[23] Yuxiang Huang and John Zelek. Motion segmentation from a moving monocular camera. 2023. 2

[24] Yuxiang Huang, Yuhao Chen, and John Zelek. Zero-shot monocular motion segmentation in the wild by combining deep learning with geometric motion model fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2733–2743, 2024. 2, 3

[25] Hossam Isack and Yuri Boykov. Energy-based geometric multi-model fitting. *International Journal of Computer Vision*, page 123–147, 2012. 2, 4

[26] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4922–4933, 2021. 3

[27] Yangbangyan Jiang, Qianqian Xu, Ke Ma, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. What to select: Pursuing consistent motion segmentation from multiple geometric

models. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 1708–1716, 2022. 3

[28] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023. 2, 5

[29] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Proc. ECCV*, 2024. 4

[30] Laurynas Karazija, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Learning segmentation from point trajectories, 2024. 2

[31] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Neural Information Processing Systems,Neural Information Processing Systems*, 2011. 6

[32] Taotao Lai, Hanzi Wang, Yan Yan, Tat-Jun Chin, and Wan-Lei Zhao. Motion segmentation via a sparsity constraint. *IEEE Transactions on Intelligent Transportation Systems*, page 973–983, 2017. 2

[33] Minhyeok Lee, Suhwan Cho, Dogyoon Lee, Chaewon Park, Jungho Lee, and Sangyoun Lee. Guided slot attention for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3807–3816, 2024. 4

[34] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *2013 IEEE International Conference on Computer Vision*, pages 2192–2199, 2013. 2, 5

[35] Long Lian, Zhirong Wu, and Stella X. Yu. Bootstrapping objectness from videos by relaxed common fate and visual grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14582–14591, 2023. 4, 5, 6, 7

[36] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 2, 5

[37] Etienne Meunier and Patrick Bouthemy. Unsupervised motion segmentation in one go: Smooth long-term model over a video, 2024. 2

[38] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. Em-driven unsupervised learning for efficient motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4462–4473, 2023. 5, 7

[39] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[40] Eslam Mohamed, Mahmoud Ewaisha, Mennatullah Siam, Hazem Rashed, Senthil Yogamani, Waleed Hamdy, Mohamed El-Dakdouky, and Ahmad El-Sallab. Monocular instance motion segmentation for autonomous driving: Kitti instancemotseg dataset and multi-task baseline. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021. 2

[41] Michal Neoral and Jan Šochman. Monocular arbitrary moving object discovery and segmentation. 3

[42] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1187–1200, 2014. 2

[43] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6): 1187–1200, 2014. 2, 5

[44] H. Opower. Multiple view geometry in computer vision. *Optics and Lasers in Engineering*, page 85–86, 2002. 3

[45] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1, 2, 3, 4

[46] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *2013 IEEE International Conference on Computer Vision*, 2013. 2

[47] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 2, 3, 5

[48] 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:1704.00675*, 2017. 2, 3, 5

[49] Mohamed Ramzy, Hazem Rashed, AhmadEl Sallab, and Senthil Yogamani. Rst-modnet: Real-time spatio-temporal moving object detection for autonomous driving. *Cornell University - arXiv,Cornell University - arXiv*, 2019. 2

[50] S Rao, R Tron, R Vidal, and Yi Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1832–1845, 2010. 2

[51] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 3, 5

[52] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15455–15464, 2021. 3

[53] Hicham Sekkati and Amar Mitiche. A variational method for the recovery of dense 3d structure from motion. *Robotics and Autonomous Systems*, 55(7):597–607, 2007. 2

[54] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. *International Jour-*

nal of Computer Vision,International Journal of Computer Vision, 2017. 2

[55] Roberto Tron and Rene Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, page 1–8, 2007. 2

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*, 2017. 4

[57] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine,IEEE Signal Processing Magazine*, 2011. 2

[58] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video, 2024. 1

[59] Andreas Wedel, Annemarie Meißner, Clemens Rabe, Uwe Franke, and Daniel Cremers. *Detection and Segmentation of Independently Moving Objects from Dense Scene Flow*, page 14–27. 2009. 2

[60] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *NeurIPS*, 2022. 5, 6, 7

[61] Junyu Xie, Weidi Xie, and Andrew Zisserman. Appearance-based refinement for object-centric motion segmentation. In *ECCV*, 2024. 2, 4, 5, 6, 7

[62] Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is sam (and flow), 2024. 2

[63] Xun Xu, Loong-Fah Cheong, and Zhuwen Li. Motion segmentation by exploiting complementary geometric models. *Cornell University - arXiv,Cornell University - arXiv*, 2018. 2

[64] Xun Xu, Loong Fah Cheong, and Zhuwen Li. Motion segmentation by exploiting complementary geometric models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2859–2867, 2018. 3

[65] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. 1

[66] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 3, 4

[67] Shichao Yang and Sebastian Scherer. Cubeslam: Monocular 3d object slam. *IEEE Transactions on Robotics*, page 925–938, 2019. 4

[68] Shu Yang, Lu Zhang, Jinqing Qi, Huchuan Lu, Shuo Wang, and Xiaoxing Zhang. Learning motion-appearance co-attention for zero-shot video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1564–1573, 2021. 3

[69] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019. 4, 5, 6, 7

[70] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018. 4

[71] Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu. Deep transport network for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8781–8790, 2021. 3

[72] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European conference on computer vision (ECCV)*, 2022. 4, 5

[73] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13066–13073, 2020. 3

[74] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13066–13073, 2020. 2