Zero-Shot Monocular Motion Segmentation in the Wild by Combining Deep Learning with Geometric Motion Model Fusion

Yuxiang Huang Yuhao Chen John Zelek
Vision and Image Processing Lab, University of Waterloo
Waterloo, ON, Canada
{yuxiang.huang, yuhao.chen1, jzelek}@uwaterloo.ca

Abstract

Detecting and segmenting moving objects from a moving monocular camera is challenging in the presence of unknown camera motion, diverse object motions and complex scene structures. Most existing methods rely on a single motion cue to perform motion segmentation, which is usually insufficient when facing different complex environments. While a few recent deep learning based methods are able to combine multiple motion cues to achieve improved accuracy, they depend heavily on vast datasets and extensive annotations, making them less adaptable to new scenarios. To address these limitations, we propose a novel monocular dense segmentation method that achieves state-of-the-art motion segmentation results in a zero-shot manner. The proposed method synergestically combines the strengths of deep learning and geometric model fusion methods by performing geometric model fusion on object proposals. Experiments show that our method achieves competitive results on several motion segmentation datasets and even surpasses some state-of-the-art supervised methods on certain benchmarks, while not being trained on any data. We also present an ablation study to show the effectiveness of combining different geometric models together for motion segmentation, highlighting the value of our geometric model fusion strategy.

1. Introduction

Motion segmentation is a fundamental problem in computer vision. It has an essential role in many applications such as action recognition, autonomous navigation, object tracking, and scene understanding in general. The objective of motion segmentation is to divide a video frame into regions segmented by common motions. Motion segmentation becomes particularly challenging when utilizing a single camera that is also moving, as this introduces issues such as degenerate motions, motion parallax, motion on the epipolar plane [18]. Existing motion segmentation methods often fails when facing these challenges since they usually rely on only a single motion cue [11, 16, 28, 35, 39, 43, 45, 49, 53], limiting their effectiveness across the diverse tapestry of real-world environments. While a few deep learning based methods are able to incorporate additional motion cues in an end-to-end manner, their reliance on large annotated datasets and the need for substantial computational resources for training limit their adaptability and application in novel environments [19, 36]. When facing these challenges, existing methods usually fail to detect the correct motion patterns and also fail to produce coherent segmentation masks for the moving objects.

In order to overcome these limitations and achieve in-the-wild monocular motion segmentation regardless of motion types and scene structures, it is necessary to have a robust and comprehensive motion model. We draw inspiration from two branches of well studied motion segmentation approaches: and point trajectory based methods and optical flow based methods. These two types of motion cues are not only complementary in nature (long-term vs short-term motion), but they can also be used to derive highly complementary geometric motion models for different motion types and scene structures. Point trajectory based methods, when analyzed using epipolar geometry, will fail if the motion is mainly on the epipolar plane or degenerate (e.g., pure forward motion), but are robust to depth variations, perspective effects and motion parallax. On the other hand, optical flow based methods do not handle these challenges well, but are robust to motions on the epipolar plane. We propose to combine these two motion cues and monocular depth information at the object level using multi-view spectral clustering, to obtain a coherent and comprehensive motion representation of the scene. By doing so, we are able to distinguish a variety of complex object motions (e.g., degenerate motions, motion parallax and non-rigid motion), even in complex scenes.

In addition to having a comprehensive geometric motion model for effectively analysis of object motions, it is also essential to obtain accurate object proposals for tracking potential moving objects throughout the video. This step is not only fundamental for accurate assessment of object motion, but also vital for generating precise and coherent segmentation masks for moving objects. To accomplish this, we leverage the strong zero-shot ability of the recent computer vision foundation models to identify, segment and track any potential moving objects throughout the video. We then calculate pairwise motion affinity scores for every object pair in the proposal, assessing how well each object-specific motion cue fits its corresponding geometric motion model. The motion affinity scores are used to construct motion affinity matrices, which can be fused by multi-view spectral clustering techniques to obtain the final clustering of objects in different motions

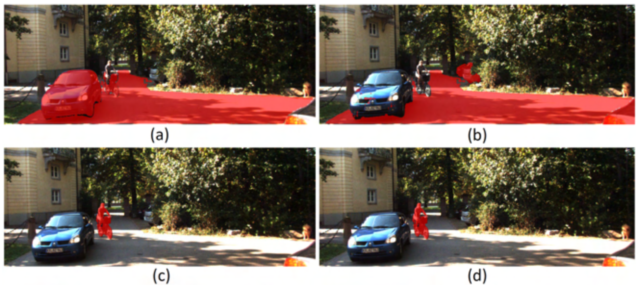


Figure 1. Motion segmentation results from the proposed method using different motion cues on a scene with motion parallax and degeneracy. Motion cues used: (a) point trajectory. (b) optical flow. (c) optical flow + depth. (d) trajectory + optical flow + depth. Using a single motion cue is insufficient to correctly segment out the moving cyclist.

Our method was evaluated on three benchmarks: DAVIS-Moving, YTVOS-Moving [11], and an extended version of the KT3DMoSeg dataset we proposed. Our method achieves competitive results on all benchmarks and even surpasses the state-of-the-art supervised method on DAVIS-Moving.

2. Related Work

Research on monocular motion segmentation has been ongoing for several decades, leading to varying interpretations of the problem among different studies. Commonly, it is defined as the process of dividing a video frame into regions that share similar motions. Alternatively, many studies also approach motion segmentation as the task of clustering predefined feature point trajectories across two or more video frames based on their distinct motions. In this paper, we focus on performing motion segmentation directly from input video frames, with the goal of segmenting entire moving objects, including those exhibiting multiple rigid motions. This approach aims at attaining a high-level understanding of the scene.

Monocular motion segmentation can be broadly divided into three distinct categories, each defined by the type of motion cues utilized. The first group consists of optical flow based methods, which rely on optical flow as their primary source of motion information [3–5, 11, 33, 35, 39, 43, 45, 46, 49, 52]. The second group includes the feature point trajectory based methods, which rely exclusively on motion information derived from manually corrected feature point trajectories throughout the video [1, 7, 12, 13, 23, 24, 30, 37, 55]. The last category comprises fusion-based methods, which combine multiple types of motion cues as well as appearance cues, to enhance the segmentation results [19, 22, 36].

2.1. Optical Flow Based Methods

Optical flow based methods can be further categorized into traditional and deep learning based methods. Traditional methods [3, 39, 45, 52] rely on optical flow masks input, and produce a pixel-wise segmentation mask indicating different motion groups. These methods usually adopt iterative optimization approaches or statistical inference techniques to estimate the motion models and motion regions simultaneously. In contrast, numerous deep learning based methods [6, 8, 14, 35, 43] use a CNN encoder to extract motion cues from optical flow and uses a decoder to produce the final segmentation. More advanced deep learning models use two CNN encoders – one to extract motion information from optical flow and the other one to extract appearance features directly from the video frame – to enhance the segmentation performance. However, deep learning methods often require a large amount of training data and do not generalize well to novel scenes.

In general, optical flow based methods perform well on scenes without strong depth variations or motion parallax. However, if the scene contains these elements (e.g. road scenes), these methods will fail to distinguish if a part of the image is moving independently or is just at a different depth from its surroundings, because the motion flow vectors projected to a 2D image from the 3D space are determined by both the depth and the screw motion of the object [34]. Additionally, strong brightness changes in the video will also adversely affect the performance of optical flow based methods since optical flow calculation is based on the brightness constancy constraint, which will be violated under strong brightness change.

2.2. Point Trajectory Based Methods

In contrast to the other two categories, point trajectory based methods produce clusters of key points that represent various motion patterns, rather than providing full dense segmentations. These techniques can be further divided into two-frame and multi-frame methods. Two frame methods [2, 12, 24] usually determine motion parameters by solving an iterative energy minimization problem of finding a certain number of geometric models (e.g., fundamental matrices) on a set of matched feature points, to minimize an energy function that evaluates the quality of the overall clustering of correspondences. Multi-frame based methods, on the other hand, usually analyze manually adjusted trajectory points from a dense optical flow tracker and often employ spectral clustering on affinity matrices. These matrices are generated through geometric model fitting [1, 20, 21, 30, 55], subspace fitting [13, 44, 48, 50], or pairwise motion affinities derived from spatio-temporal motion cues and appearance cues [7, 37].

The efficacy of point trajectory based methods is heavily influenced by the chosen motion model and the precision of point correspondences. There is not a single motion model that can capture motion similarities across all types of motions. In search of a better motion model, [1] uses trifocal tensor to analyze point trajectories. Trifocal tensor is more robust to noises and is able to distinguish motions on the epipolar plane, but it is harder to optimize and prone to failure when the three cameras are close to being colinear [18], which can often happen on road scenes. [26, 55] proposed geometric model fusion techniques to combine different geometric models, but they still fail to produce coherent and consistent segmentations on complex scenes. Moreover, most existing methods depend on manually refined point correspondences and struggle to effectively manage outliers.

2.3. Fusion Based Methods

Recent research in motion segmentation have introduced several innovative approaches that leverage a combination of motion cues for improved motion segmentation accuracy. Notably, [36] integrates optical flow masks and monocular depth maps through a fusion module in their neural network, facilitating end-to-end training. By adopting a semi-supervised training strategy, this approach has set new benchmarks in monocular dense motion segmentation across various datasets. Another study [19] explored the impact of utilizing different combinations of motion cues, such as optical flow, depth map, and scene flow, on motion segmentation performance, achieving state-of-the-art results on the KITTI and DAVIS datasets. However, this approach is fully supervised, which requires an extensive amount of training data and computational power. Furthermore, its ability to generalize across a wider range of benchmarks remains unverified. In our earlier work [22], we introduced an interpretable geometric model that merges optical flow with monocular depth maps for zero-shot motion segmentation. Despite these efforts, a noticeable performance gap persists between our method and the state-of-the-art supervised or semi-supervised techniques, showing the insufficiency of relying solely on a single geometric model to achieve optimal results in motion segmentation.

Despite these recent research, no existing method has yet to combine the two complementary and most commonly explored motion cues in motion segmentation: point trajectory and optical flow. This paper seeks to fill this gap, demonstrating how combining these motion cues with well-crafted geometric motion models can lead to state-of-the-art zero-shot monocular motion segmentation.

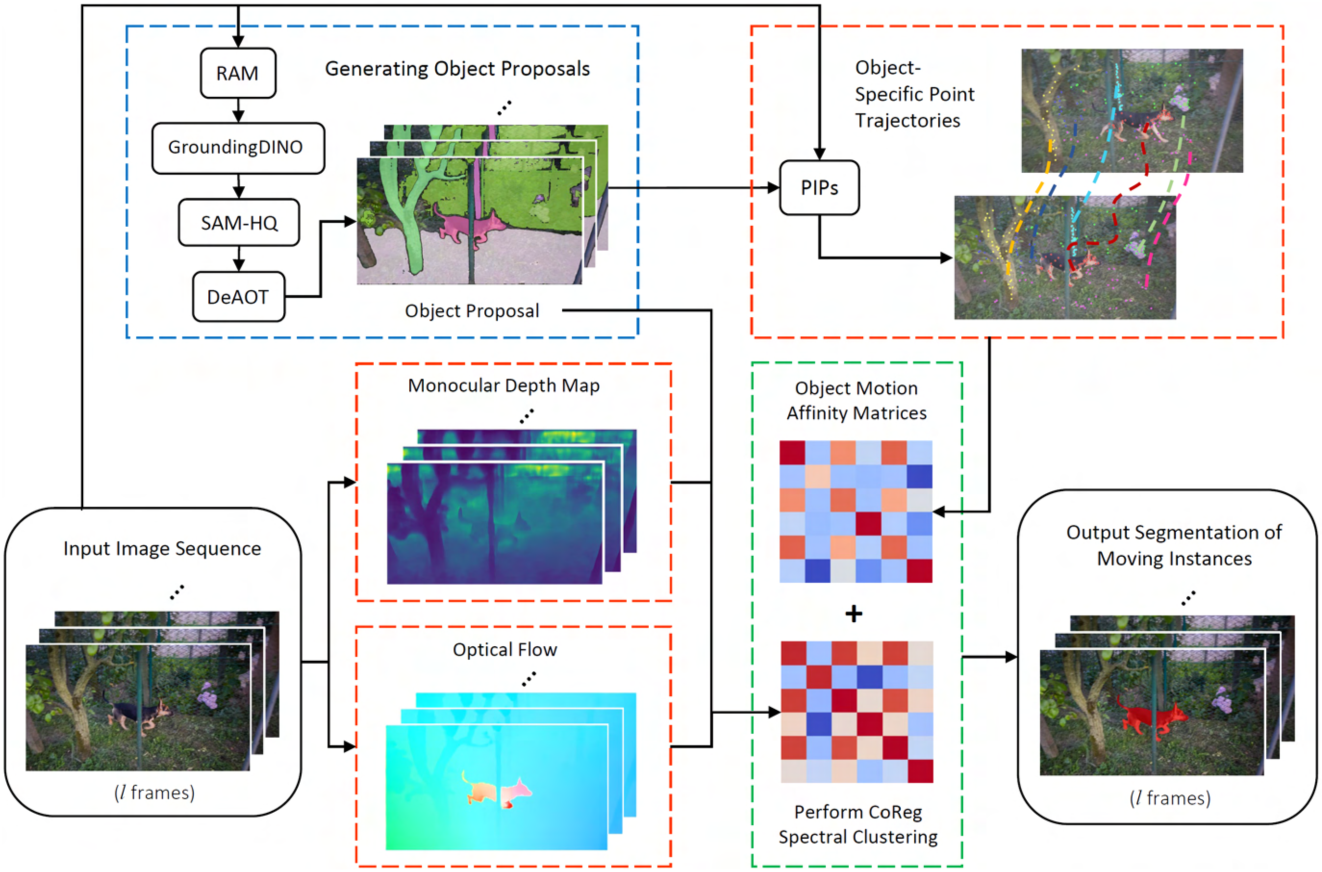
3. Methodology

We propose a zero-shot monocular motion segmentation approach that uses both object appearance information and a combination of epipolar geometry and optical flow based geometric motion models to perform in-the-wild motion segmentation without any assumptions of the motion or the scene that may appear in the video.

Our segmentation pipeline begins by identifying initial segmentation of the background and common objects within the scene through foundational models, followed by continuous tracking of these objects across the video sequence. For every object in each frame, we gather a collection of object-specific trajectory points, an optical flow mask, and a monocular depth map. Subsequently, we construct two distinct geometric motion models for each scene object: one via fundamental matrix fitting using point trajectories and the other via fitting optical flow and a depth map to our proposed parametric equations. By fitting each object's motion models on every other object and analysing the residuals of the model fitting, we are able to derive two pairwise affinity scores between every pair of objects, from which we can construct two motion affinity matrices for the two types of motion models respectively. Lastly, we fuse the two affinity matrices using co-regularized multi-view spectral clustering to obtain the final segmentation. Figure 2 shows a diagram of the motion segmentation pipeline.

3.1. Generating Object Proposals

In order to identify all motions in a video sequence at object level, we use the same method as proposed in [22] to identify, segment and track each prominent object across the video. This is accomplished by integrating foundation models for object recognition (RAM) [58], detection (Grounding DINO model) [31], segmentation (SAM-HQ) [27], and tracking (DeAOT) [57]. This video preprocessing pipeline for automatic object proposal generation is inspired by and improved upon Segment and Track Anything (SAMTrack) [9]. Comparing to SAMTrack, our video preprocessing module combines these foundation models to segment and track objects automatically, bypassing the need for manual text prompts by initiating our pipeline with the Recognize Anything Model to automatically detect common objects in the initial video frame. Our object proposal generation pipeline involves: 1) Automatically identifying common



RAM
Generating Object Proposals
GroundingDINO
SAM-HQ
DeAOT
Object Proposal
Object-Specific Point Trajectories
PIPs
Monocular Depth Map
Object Motion Affinity Matrices
Input Image Sequence
(l frames)
Optical Flow
+
Perform CoReg Spectral Clustering
Output Segmentation of Moving Instances
(l frames)

Figure 2. Our Motion Segmentation Pipeline. Our method can be summarized to three main steps: 1) given a sequence of video frames, we produce an object proposal by automatically detecting, segmenting and tracking common objects in the video. 2) we compute object-specific point trajectories, optical flow and monocular depth maps for every frame. 3) we compute pairwise object motion similarity scores using two motion models (one based on point trajectories and the other based on optical flow and depth map), and use them to construct two motion affinity matrices. The two matrices are fused using multi-view spectral clustering to cluster objects into different motion groups.

objects in the video's first frame using the Recognize Anything Model; 2) Generating object bounding boxes with the Grounding DINO model; 3) Producing instance segmentation masks for the initial frame via the SAM-HQ model, applying non-max suppression to refine the results; 4) Employing the DeAOT tracker to track each object's segmentation mask throughout the video. To accommodate new objects appearing mid-sequence, we segment the video into sections of l frames, repeating the above process for each section. The choice of l varies depending on the video's dynamics and the frequency of new objects entering the scene. Videos with higher dynamics and frequent entry of new objects mid-sequence are better suited to a reduced segment length l.

3.2. Object-Specific Motion Cues

Once we have an object proposal for every frame of the video, we will then obtain object-specific motion cues for every object in the object proposal. We propose to use point trajectories, optical flow and monocular depth map automatically generated by off-the-shelf networks as motion cues, in order to model objects' motions in two complementary ways.

3.2.1 Object-Specific Point Trajectories

A set of sparse point trajectories is generated for every object using PIPs [17]. PIPs is a state-of-the-art point tracker which tracks individual pixels given their initial locations in a video frame. A mixture of Shi-Tomasi [25] and K-Medoids [40] sampling method is used to obtain the initial pixels from each object as it showed good experimental results from previous works in similar tasks [42]. These tracked pixels can be used as object-specific feature points to fit fundamental matrices for every object in frame pairs to describe their motions. One limitation of PIPs is that does not handle occlusion well if the tracked video is more than 8 frames. To overcome this issue, we check for every point if it is inside its corresponding object's mask area every 8 frames. If not, we remove that point and sample a new point inside the object's mask. We also remove any point that is near the edge of the frame since the tracking accuracy of PIPs drops significantly in this case.

3.2.2 Object-Specific Optical Flow and Depth Map

We also generate a dense optical flow mask and a monocular depth map for every frame, from which we can extract object-specific optical flow vectors and depth maps. We use a state-of-the-art optical flow estimator [47] to obtain optical flow, and a state-of-the-art monocular depth estimator, DINOv2 [38], to extract the depth maps. We use monocular depth estimation to estimate the scene depth from a single frame since our goal is to perform motion segmentation from a moving monocular camera. DINOv2 outputs a relative depth map, which is sufficient for our application. Our experiment shows improved results when both optical flow and depth map are used to compute the motion model, comparing to only optical flow. We show how a depth map can be used to improve the motion model based solely on optical flow in the next section.

3.3. Geometric Motion Model Fitting

After obtaining object-specific point trajectories, optical flow vectors and depth maps, for each frame pair, we compute two geometric motion models of each object based on epipolar geometry and optical flow respectively, to model its motion throughout the video. To compute the epipolar geometry based motion models using point trajectories, we compute a fundamental matrix of each object between every *f* frames by solving $p'TFp = 0$ using the eight-point algorithm with RANSAC [15], where p and p' are the normalized 2D homogeneous coordinates of the same tracked point in the two frames. If a degenerate case is encountered for the fundamental matrix, we do not use it.

For the optical flow and depth based motion model, we use the same motion model proposed in our earlier work [22]. We refine the Longuet-Higgins and Pruzdny model equation [32] to address rigid object motion, adapting it to include depth information without needing exact pixel depth, a common limitation in practice. Instead of relying on the original model, which is impractical due to unknown absolute pixel depth, we propose a linearized version incorporating relative depth from DINOv2, making it more applicable to real-world scenarios with varying depths. This approach, while using both optical flow and depth data, simplifies the motion model to the following linear equations:

$$\begin{aligned} u &= a + b\frac{1}{z} - c\frac{x}{z} - dy + ex^2 - fxy \\ v &= g + h\frac{1}{z} - c\frac{y}{z} - dx + exy + fy^2 \end{aligned} \tag{1}$$

This motion model aims to cluster different motions rather than calculate exact screw motions, sidestepping scale uncertainties and making it theoretically sound without requiring specific camera intrinsics. For consistency reasons, we still refer to this motion model as the ”optical flow motion model”, although it uses both optical flow vectors and pixel depth maps.

3.4. Constructing Motion Affinity Matrices

After all fundamental matrices and optical flow motion models are computed, each object will have a fundamental matrix between every p frames and an optical flow motion model between every two frames. By fitting every object's trajectory points, optical flow vectors and depth maps to every other object's fundamental matrix and optical flow motion model on the same frame pair, we can obtain the residuals of every object to all other objects' motion models respectively. We use Sampson distance [18] as the residual for the fundamental matrix and mean squared error for the optical flow motion model. Assuming there are *k* objects in the scene, for the *i*-th object at the *m*-th frame pair, we obtain the following residual vectors under the fundamental matrix and optical flow motion models:

$$\boldsymbol{r}_{o_i}^{m} = [r_{o_{i,1}}^{m}, r_{o_{i,2}}^{m}, ..., r_{o_{i,k}}^{m}],$$

$$\boldsymbol{r}_{f_i}^{m} = [r_{f_{i,1}}^{m}, r_{f_{i,2}}^{m}, ..., r_{f_{i,k}}^{m}]$$

where $r_{o_{i,k}}^{m}$ is the mean residual for fitting the parametric motion model of object i on the optical flow vectors of object k between frames m and $m+1$, and $r_{f_{i,k}}^{m}$ is the mean Sampson error for fitting the fundamental matrix of object i on the trajectory points of object k between frames m and $m+p$. We construct two affinity matrices encapsulating the pairwise motion affinities between each pair of objects using a modified version of ordered residual kernal (ORK) [10]. Specifically, for each object, we sort its residual vectors in ascending order and define a threshold to select the smallest *t*-th residual as inliers. We define $\boldsymbol{c}_i = \{0, max(t - n_i, 0)\}^K$ as an inlier score vector, where n_i is the rank of object k in the residual vector of object i, penalizing different inlier distributions between objects. The pairwise motion affinity between objects i and j can thus be computed as $\boldsymbol{a}_{ij} = \boldsymbol{c}_i^\intercal \boldsymbol{c}_j$, which denotes a weighted co-occurrence score between two objects as inliers of all motion models. Our proposed weighted ORK is robust to outliers and makes the affinity matrix more adaptive to different scenes by reducing the need to set scene specific inlier thresholds.

3.5. Co-Regularized Multi-view Spectral Clustering

After constructing the affinity matrices, we normalize them using row normalization [51] and adapt co-regularized

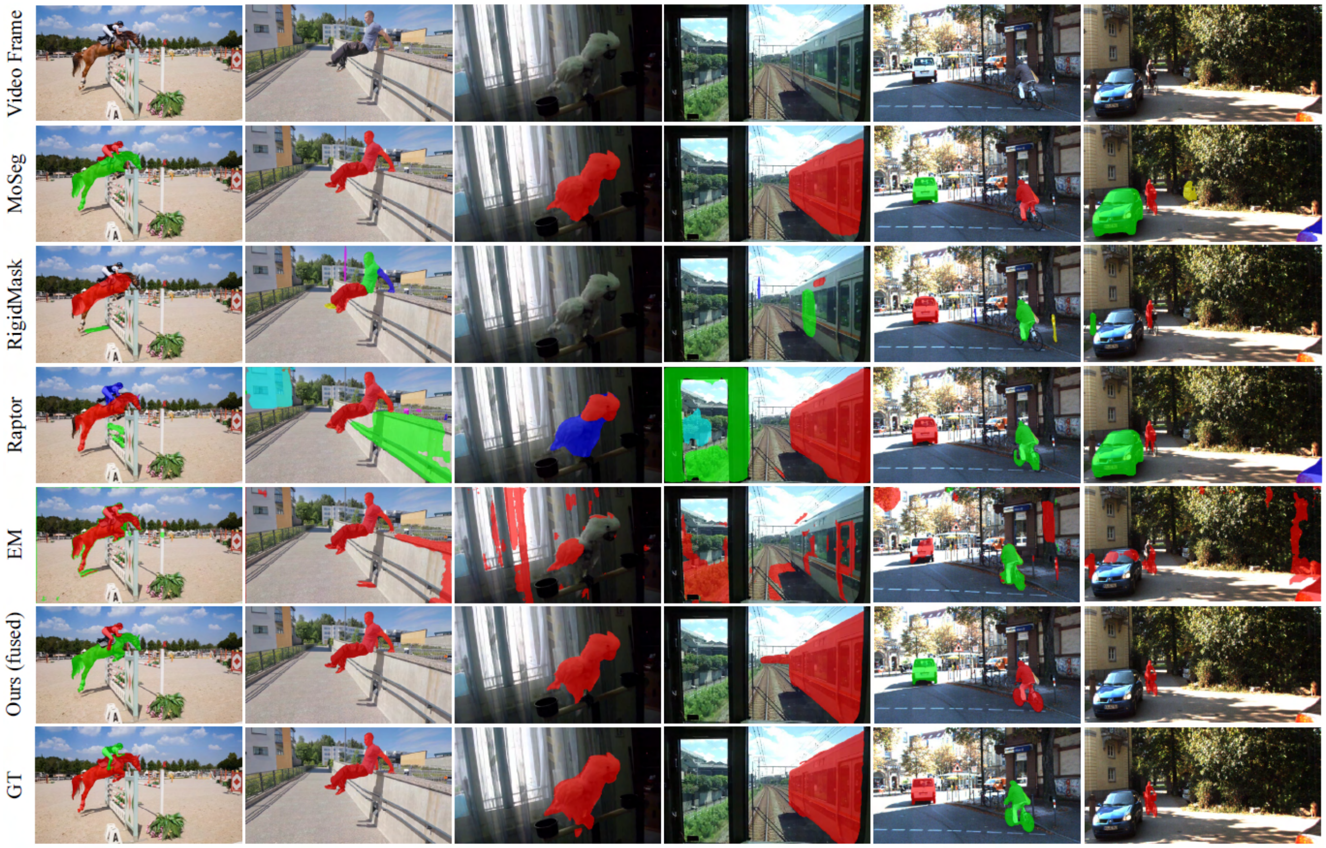


Figure 3. Qualitative results of different methods on DAVIS-Moving (row 1, 2), YTVOS-Moving (row 3, 4) and the extended KT3DMoSeg (row 5, 6) datasets. MoSeg often mistakenly label static objects as dynamic when there is degenerate camera motion. RigidMask fails to detect or coherently segment objects with non-rigid motions. Similarly, Raptor also has these problems, although to a lesser extent overall. Our method, despite being zero-shot, performs well when facing these challenges.

multi-view spectral clustering [29] to fuse the two affinity matrices together. With the number of motion groups in the scene given as an input, we are able to obtain the final clustering of moving objects. Co-regularized multi-view spectral clustering uses an regularization term to encourage consensus between different views and is shown to perform well on fusing multiple geometric models for a consistent representation of motion information [55].

4. Experiments

Our method is tested on three benchmarks: DAVIS-Moving, YTVOS-Moving and the extended KT3DMoSeg. We first briefly introduce these datasets, then show both quantitative and qualitative comparisons between our method and other state-of-the-art methods. Lastly, we present an ablation study to compare the performance of each individual motion models and the fused motion model.

4.1. Datasets

DAVIS-Moving and YTVOS-Moving are both proposed by [11] as datasets for generic instance motion detection and segmentation. DAVIS-Moving and YTVOS-Moving are subsets of the DAVIS 17 dataset [41] and the YTVOS dataset [54], where all moving instances in the video sequence are labeled and no static objects are labeled. These two recently proposed datasets are very challenging due to their diverse object classes, occlusions and non-rigid motions.

In addition to these two datasets, we also evaluate our method on an extended version of the KT3DMoSeg dataset. The original KT3DMoSeg dataset [55] is designed to test point trajectory based motion segmentation methods on complex road scenes. It contains manually corrected point trajectories on selected moving instances in road scenes and includes significant degenerate motions and depth variation. In order to test the performance of our method in such environments, we extend the KT3DMoSeg dataset by adding a pixel-level segmentation mask to every moving instance in the scene. We refer to this extended dataset as the KT3DInsMoSeg dataset in the following sections.

Exp.	Method	Training	DAVIS-Moving			YTVOS-Moving			KT3DInsMoSeg		
			Pu	**Ru**	**Fu**	**Pu**	**Ru**	**Fu**	**Pu**	**Ru**	**Fu**
a	MoSeg [11]	Supervised	**78.30**	78.80	78.10	**74.50**	66.40	**66.38**	63.73	78.24	66.85
a	Raptor [36]	Supervised Features	75.90	79.67	75.93	64.43	60.94	60.35	71.52	**88.27**	**75.82**
a	RigidMask [56]		59.03	49.89	50.01	29.88	17.88	18.70	65.14	83.34	70.91
a	EM [33]	Unsupervised	58.42	83.48	64.24	44.52	40.33	37.12	42.85	58.71	44.03
a+b	**Ours (fused)**	Zero-Shot (no training)	78.27	81.58	**79.40**	64.12	61.10	60.62	**72.93**	71.02	71.89
b	Ours (OC+depth)		71.53	75.66	73.18	63.54	58.94	56.06	48.04	61.54	49.26
b	Ours (OC)		58.25	59.22	57.08	61.79	54.64	53.74	36.44	39.97	34.78
b	Ours (trajs)		65.99	75.51	68.47	54.67	52.92	50.05	42.31	73.66	45.24
b	Ours (base)		43.17	**86.24**	52.12	48.49	**73.01**	50.82	38.97	70.97	43.37

Table 1. Performance of our method and state-of-the-art motion segmentation methods (**Exp.** a) on the DAVIS-Moving, YTVOS-Moving validation datasets and the KT3DInsMoSeg dataset, as well as ablation study results (**Exp.** b). The best result for each metric is in bold and the second best result is in underscore. Our method overall performs the best on DAVIS-Moving and second best on both YTVOS and KT3DInsMoSeg, despite not being trained on any data. Our method also significantly surpasses the state-of-the-art unsupervised motion segmentation method [33].

4.2. Results and Discussion

Our method's performance is evaluated using *precision* (Pu), *recall* (Ru), and *F-measure* (Fu) proposed in [11] which penalizes false positives. The *F-measure* combines both *precision* and *recall* and indicates the method's overall performance. Table 1 shows quantitative results of our method and other state-of-the-art methods on the three benchmarks. Despite no training, our approach excels on the DAVIS-Moving dataset, outperforming fully-supervised methods, ranks second on the YTVOS-Moving dataset, closely surpassing Raptor [36], and secures a similar position on the KT3DInsMoSeg dataset, trailing only behind Raptor. Our method also significantly surpasses EM [33], which is the state-of-the-art unsupervised multi-label motion segmentation method.

We also qualitatively compare our method with these methods and show the results in 3. Results indicate our method's superiority in identifying static and moving objects across various scenes, notably in complex scenarios where other methods fail, such as in scenes with degenerate motions or complex object contours. Our technique demonstrates robust performance across all datasets, showing its effectiveness in accurately grouping motions and outperforming existing methods in challenging conditions.

One primary limitation of the proposed method is its inference speed. Despite being a zero-shot approach that requires no training, the method integrates multiple computer vision foundation models, as well as the neural networks for feature point tracking and optical flow estimation. Consequently, such integration significantly slows the method's processing speed, making it only suitable to be applied on pre-recorded video s. Another limitation is the requirement for a known ground truth number of motions in the scene to achieve optimal results, inherent to the use of spectral clustering. Although this issue can be mitigated by employing various model selection methods [21, 51], such adjustments typically result in a slight degradation of performance.

4.3. Ablation Study

We present both quantitative (Table 1, Exp. b) and qualitative (Figure. 4) comparisons between different individual motions models and the fused motion model for their performances on the three benchmarks.

We found that on both DAVIS-Moving and KT3DInsMoSeg datasets, our model fusion technique (fused) is able to significantly boost the Fu score comparing to using only a single model, while on YTVOS-Moving, the Fu score only had a relatively small increase. Upon further inspection, we discovered this could be attributed to some motion labels in the YTVOS-Moving dataset actually being mostly static throughout the video sequence. Since our method clusters moving objects purely using motion cues, it groups these objects together with the background as expected. Additionally, the YTVOS-Moving dataset also contains videos with significant camera zooming, which violates the geometric assumptions of both our motion models. Our motion model fusion technique is able to achieve better results than any single motion model on all three datasets, showing its effectiveness.

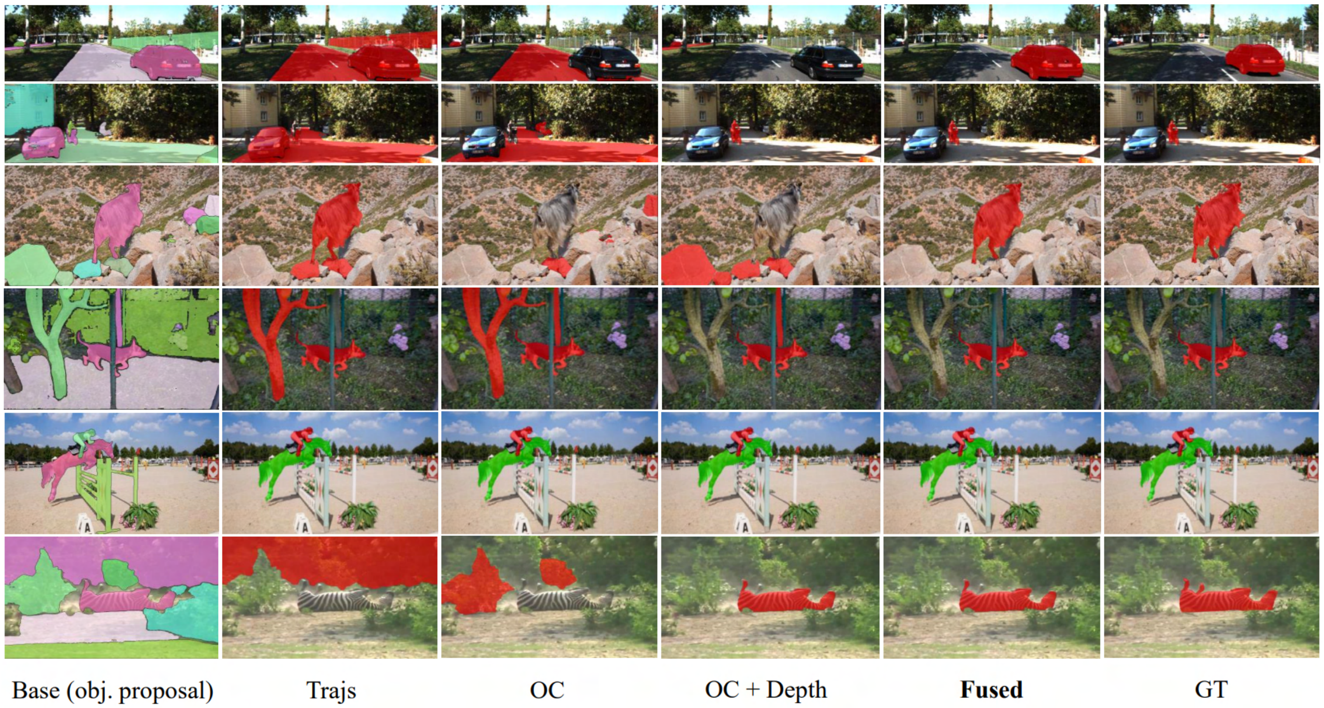


Figure 4. Qualitative comparison of different motion models on different scenes. Pure optical flow based motion model (OC) suffers on scenes with objects at varying depths. Combining optical flow with depth information (OC + Depth) only alleviates this problem to some extent. Pure point trajectory based motion model (Trajs) suffers from motions near the epipolar plane and inaccurate trajectory estimation. Motion model fusion solves these problem by combining the advantages of both motion models and outperforms any single model.

We also show the motion segmentation performance of our pipeline under conditions where only partial motion cues are used. Specifically, we present results obtained from two different types motion affinity matrices, which are computed using two different motion models: one solely based on the optical flow motion model (OC), and another that combines optical flow with monocular depth information (OC + depth). The optical flow based motion model is obtained from [33], which is a state-of-the-art unsupervised method using only optical flow as input. The motion model combining optical flow and depth is proposed by [22], which is a direct improvement on [33]. Results show that the motion model based on a combination of optical flow and depth (OC + depth) outperforms OC by a large margin in all three metrics on both DAVIS-Moving and KT3DInsMoSeg, while having similar results on YTVOS-Moving.

Both Point trajectory based (trajs) and optical flow based motion models perform poorly on the KT3DInsMoSeg dataset, potentially due to significant motion degeneracy (e.g., forward motion) and depth variations on road scenes. Incorporating depth information in this case proves to be an effective way to reduce motion ambiguity for the optical flow based motion model, boosting its F-score from 34.78% to 49.26%. Fusing the combined (OC + depth) motion model with the epipolar geometry based point trajectory motion model significantly enhances the performance in this case.

5. Conclusion and Future Work

We propose the first zero-shot monocular motion segmentation approach that achieves state-of-the-art performance. Our method combines the advantages of both deep learning and multiple geometric approaches, resulting in a zero-shot motion segmentation approach that performs geometric motion model fusion on object proposals. We compare the performances of the fused motion model and each individual motion model, and observe a significant performance improvement for the fused motion model, showing the effectiveness of the proposed geometric motion model fusion technique. Even though our method is zero-shot, experiments show that our method is better than many state-of-the-art methods and highly competitive with others.

Future research could pursue two promising directions: First, the integration of additional motion models, such as the trifocal tensor [18], may further improve the motion segmentation performance. Second, developing methods to effectively incorporate both types of motion affinity measures into the loss function could enable end-to-end, self-supervised training of a motion segmentation network, potentially achieving substantial improvement in inference speed.

References

[1] Federica Arrigoni, Luca Magri, and Tomas Pajdla. On the Usage of the Trifocal Tensor in Motion Segmentation. In *Computer Vision – ECCV 2020*, pages 514–530. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science. 2, 3

[2] Daniel Barath and Jiri Matas. Progressive-X: Efficient, Anytime, Multi-Model Fitting Algorithm. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3779–3787, Seoul, Korea (South), 2019. IEEE. 2

[3] Pia Bideau and Erik Learned-Miller. It's Moving! A Probabilistic Model for Causal Motion Segmentation in Moving Camera Videos. In *Computer Vision – ECCV 2016*, pages 433–449. Springer International Publishing, Cham, 2016. Series Title: Lecture Notes in Computer Science. 2

[4] Pia Bideau, Aruni RoyChowdhury, Rakesh R. Menon, and Erik Learned-Miller. The Best of Both Worlds: Combining CNNs and Geometric Constraints for Hierarchical Motion Segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 508–517, Salt Lake City, UT, USA, 2018. IEEE.

[5] Pia Bideau, Rakesh R. Menon, and Erik Learned-Miller. MoA-Net: Self-supervised Motion Segmentation. In *Computer Vision – ECCV 2018 Workshops*, pages 715–730. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science. 2

[6] Markus Bosch. Deep Learning for Robust Motion Segmentation with Non-Static Cameras, 2021. arXiv:2102.10929 [cs]. 2

[7] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of the 11th European conference on Computer vision: Part V*, pages 282–295, Berlin, Heidelberg, 2010. Springer-Verlag. 2, 3

[8] 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)*, pages 5587–5596, Long Beach, CA, USA, 2019. IEEE. 2

[9] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and Track Anything, 2023. arXiv:2305.06558 [cs]. 3

[10] Tat-jun Chin, Hanzi Wang, and David Suter. The Ordered Residual Kernel for Robust Motion Subspace Clustering. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2009. 5

[11] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards Segmenting Anything That Moves. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1493–1502, Seoul, Korea (South), 2019. IEEE. 1, 2, 6, 7

[12] 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*, pages 2173–2180, 2010. ISSN: 1063-6919. 2

[13] E. Elhamifar and R. Vidal. Sparse Subspace Clustering: Algorithm, Theory, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. 2, 3

[14] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. EpO-Net: Exploiting Geometric Constraints on Dense Trajectories for Motion Saliency. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1873–1882, Snowmass Village, CO, USA, 2020. IEEE. 2

[15] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[16] K. Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1853, Providence, RI, 2012. IEEE. 1

[17] Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. In *Computer Vision – ECCV 2022*, pages 59–75. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science. 4

[18] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1, 3, 5, 8

[19] Christian Homeyer and Christoph Schnorr. On Moving Object Segmentation from Monocular Video with Transformers. 1, 2, 3

[20] Yuxiang Huang and John Zelek. Motion Segmentation from a Moving Monocular Camera. In *IROS 2023 Workshop on Robotic Perception and Mapping: Frontier Vision and Learning Techniques*. arXiv, 2023. arXiv:2309.13772 [cs]. 3

[21] Yuxiang Huang and John Zelek. A Unified Model Selection Technique for Spectral Clustering Based Motion Segmentation. *Journal of Computational Vision and Imaging Systems*, 9(1), 2023. 3, 7

[22] Yuxiang Huang, Yuhao Chen, and John Zelek. Dense Monocular Motion Segmentation Using Optical Flow and Pseudo Depth Map: A Zero-Shot Approach. In *21st Conference on Robots and Vision (CRV)*, Guelph, ON, Canada, 2024. IEEE. 2, 3, 5, 8

[23] David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Thomas Brox, and Jitendra Malik. Object Segmentation by Long Term Analysis of Point Trajectories. In *Computer Vision – ECCV 2010*, pages 282–295. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. Series Title: Lecture Notes in Computer Science. 2

[24] Hossam Isack and Yuri Boykov. Energy-Based Geometric Multi-model Fitting. *International Journal of Computer Vision*, 97(2):123–147, 2012. 2

[25] Jianbo Shi and Tomasi. Good features to track. In *Proceedings of IEEE Conference on Computer Vision and Pattern*

Recognition CVPR-94, pages 593–600, Seattle, WA, USA, 1994. IEEE Comput. Soc. Press. 4

[26] 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*, 35(2):1708–1716, 2021. Number: 2. 3

[27] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment Anything in High Quality, 2023. arXiv:2306.01567 [cs]. 3

[28] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion Trajectory Segmentation via Minimum Cost Multicuts. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3271–3279, 2015. ISSN: 2380-7504. 1

[29] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized Multi-view Spectral Clustering. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2011. 6

[30] 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*, 18 (4):973–983, 2017. Conference Name: IEEE Transactions on Intelligent Transportation Systems. 2, 3

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, 2023. arXiv:2303.05499 [cs]. 3

[32] H. C. Longuet-Higgins and K. Prazdny. The Interpretation of a Moving Retinal Image. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 208(1173):385–397, 1980. Publisher: The Royal Society. 5

[33] 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. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 2, 7, 8

[34] Amar Mitiche and J.K. Aggarwal. *Computer Vision Analysis of Image Motion by Variational Methods*. Springer International Publishing, Cham, 2014. 2

[35] 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)*, pages 114–121, Nagoya, Japan, 2021. IEEE Press. 1, 2

[36] Michal Neoral. Monocular Arbitrary Moving Object Discovery and Segmentation. 1, 2, 3, 7

[37] 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. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 2, 3

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2023. 5

[39] Anestis Papazoglou and Vittorio Ferrari. V.: Fast object segmentation in unconstrained video. In *In: ICCV (2013*. 1, 2

[40] 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. 4

[41] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation, 2018. arXiv:1704.00675 [cs]. 6

[42] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment Anything Meets Point Tracking, 2023. 4

[43] Mohamed Ramzy, Hazem Rashed, Ahmad El Sallab, and Senthil Yogamani. RST-MODNet: Real-time Spatio-temporal Moving Object Detection for Autonomous Driving, 2019. arXiv:1912.00438 [cs, stat] version: 1. 1, 2

[44] Shankar Rao, Roberto Tron, Rene 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. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 3

[45] 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. 1, 2

[46] Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand, and Ahmad El-Sallab. MODNet: Motion and Appearance based Moving Object Detection Network for Autonomous Driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2859–2864, 2018. ISSN: 2153-0017. 2

[47] Deqing Sun, Charles Herrmann, Fitsum Reda, Michael Rubinstein, David J. Fleet, and William T. Freeman. Disentangling Architecture and Training for Optical Flow. In *Computer Vision – ECCV 2022*, pages 165–182, Cham, 2022. Springer Nature Switzerland. 5

[48] 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*, pages 1–8, 2007. ISSN: 1063-6919. 3

[49] Johan Vertens, Abhinav Valada, and Wolfram Burgard. SMSnet: Semantic motion segmentation using deep convolutional neural networks. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 582–589, 2017. ISSN: 2153-0866. 1, 2

[50] Rene Vidal. Subspace Clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011. 3

[51] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. 5, 7

[52] Andreas Wedel, Annemarie Meißner, Clemens Rabe, Uwe Franke, and Daniel Cremers. Detection and Segmentation of Independently Moving Objects from Dense Scene Flow. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 14–27. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. Series Title: Lecture Notes in Computer Science. 2

[53] Andreas Wedel, Annemarie Meißner, Clemens Rabe, Uwe Franke, and Daniel Cremers. Detection and Segmentation of Independently Moving Objects from Dense Scene Flow. pages 14–27, 2009. 1

[54] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In *Computer Vision – ECCV 2018*, pages 603–619. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science. 6

[55] Xun Xu, Loong Fah Cheong, and Zhuwen Li. Motion Segmentation by Exploiting Complementary Geometric Models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2867, Salt Lake City, UT, USA, 2018. IEEE. 2, 3, 6

[56] Gengshan Yang and Deva Ramanan. Learning to Segment Rigid Motions from Two Frames. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1266–1275, Nashville, TN, USA, 2021. IEEE. 7

[57] Zongxin Yang and Yi Yang. Decoupling Features in Hierarchical Propagation for Video Object Segmentation. *Advances in Neural Information Processing Systems*, 35: 36324–36336, 2022. 3

[58] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize Anything: A Strong Image Tagging Model, 2023. arXiv:2306.03514 [cs]. 3