On Moving Object Segmentation from Monocular Video with Transformers

Christian Homeyer
Robert Bosch GmbH, Corporate Research, Computer Vision Lab Hildesheim, Germany
Image and Pattern Analysis Group, Heidelberg University, Germany
homeyer@math.uni-heidelberg.de

Christoph Schnörr
Image and Pattern Analysis Group, Heidelberg University, Germany
schnoerr@math.uni-heidelberg.de

Abstract

Moving object detection and segmentation from a single moving camera is a challenging task, requiring an understanding of recognition, motion and 3D geometry. Combining both recognition and reconstruction boils down to a fusion problem, where appearance and motion features need to be combined for classification and segmentation.
In this paper, we present a novel fusion architecture for monocular motion segmentation - M^3Former, which leverages the strong performance of transformers for segmentation and multi-modal fusion. As reconstructing motion from monocular video is ill-posed, we systematically analyze different 2D and 3D motion representations for this problem and their importance for segmentation performance. Finally, we analyze the effect of training data and show that diverse datasets are required to achieve SotA performance on Kitti and Davis. Code will be released upon publication.

1. Introduction

Interaction in a dynamic world requires reasoning about your surroundings and other dynamic agents. Motion segmentation plays a crucial part in autonomous perception systems, as we need this information for higher-level planning and navigation. It has exciting applications in downstream tasks such as e.g. Neural Scene Synthesis [37] or Simultaneous Localization and Mapping (SLAM) [82]. Humans and animals can effortlessly perceive even completely unknown objects when observing them moving. This is in stark contrast to common image detectors [13], which are trained on large-scale datasets and are dependent on their respective finite label spaces. Combining motion and appearance data can resolve this issue and create generic object detectors, that generalize better across datasets [18, 44].

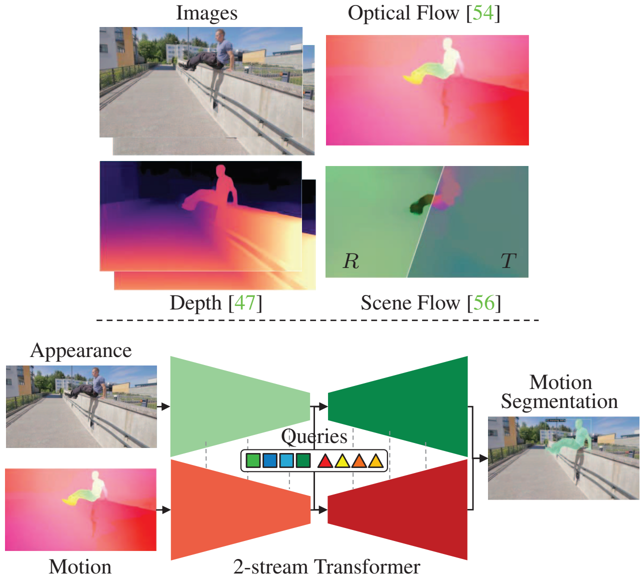

Figure 1: Our **M**ulti-**M**odal **M**ask2Former (**M**^{3}Former) framework for motion segmentation. Based on a monocular video, we compute a reconstruction based on frozen expert models [54, 56, 47]. This allows us to create (pseudo-) multi-modal data. We perform motion segmentation as a top-down fusion task with a segmentation transformer. We experiment both with 2D and 3D motion as input to our model.

These findings align with the two-stream hypothesis in Neuroscience [23], which states that both appearance and motion are vital to biological visual systems. Motion segmentation can therefore be considered a multi-modal fusion problem. In this paper, we present a novel two-stream fusion architecture for motion segmentation. We combine both appearance and motion features in a transformer architecture [13].

We call our framework **M**ulti-**M**odal **M**ask2Former (**M**^{3}Former), since we combine information from multiple modalities with masked attention. Since monocular video provides only a single modality stream, we make use of frozen expert models [47, 54, 56] for computing different motion representations, see Figure 1. Our contributions are fourfold:

- We design a novel two-stream architecture with Encoder and Decoder. We analyze the performance of different fusion strategies within this framework.
- We systematically analyze the effect of different motion representations (Optical Flow, Scene Flow, higher-dimensional embeddings) from previous work within our framework.
- We empirically showcase the effect of diverse training data. Balancing different sources of motion patterns and semantic classes is crucial for strong performance on real-life video.
- We introduce a very simple augmentation technique for better multi-modal alignment. By introducing neg. examples with no motion information, we force the network to not over-rely on appearance data alone.

1.1. Problem Statement

Given a video $\{I_1,\ I_2,\ \ldots,\ I_N\}$ from a single camera, we want to detect and segment *generic independently moving objects*. An *object* is defined as a spatially connected group of pixels, belonging to the same semantic class. All labels are merged into a single "object", since only the motion state matters. Detectors only see a finite number of classes during training. *Generic* object detection assumes an inbalance between the set of training and test class labels. We want to identify any moving object, even if we have never seen the class during training. An object is defined as *independently moving* when its apparent motion is not due to camera ego-motion. The object is still considered moving when only a part is in motion, e.g. when a person moves an arm, then the whole person should be segmented.

2. Related Work

Segmenting objects based on their motion is a long standing problem in Computer Vision with a rich history [17, 28, 58, 59, 60, 52, 61, 65, 72, 80, 9, 45, 5, 67] dating back to the early 90's.

Spatio-temporal Grouping and Geometric Modeling. Traditional approaches treat the problem as a spatio-temporal grouping problem, where similar 3D motions are clustered together [58, 52, 65, 64, 73, 9, 45, 21, 5, 62, 72]. However, they focus on theoretical analysis with perfect input data, work on simplistic scenes and/or use sparse point trajectories.

A dominant line of work focuses on segmentation from two-frame optical flow, either by devising handcrafted geometric constraints [5, 57], e.g. motion angle and plane plus parallax (P+P) [51], or by learning directly from motion data [6, 78, 35, 74, 69]. Such approaches are affected by noisy inputs and cannot deal with degenerate cases like coplanar-colinear motion [80] and camera motion degeneracy [59]. Similar to us, [40] uses two RGB-D frames as input data and use a CNN to separate static background and dynamic foreground. However, they focus on high-quality depth maps and model motion with 2D optical flow and camera poses. In order to deal with all motion cases and have a generic approach, [76] formulates extensive criteria beyond 2D motion. This requires a depth prior [47] and additional specialized neural network modules [75, 8]. Our approach is indifferent towards geometric modeling: We analyze the importance of motion models in Section 4 by ablating different representations common in the literature. We will later see, that the effect on the downstream segmentation task is highly dependent on the datasets involved and the underlying quality of the geometric model inputs. Interestingly, weaknesses in geometric modelling can be compensated with local and global image information very effectively.

Learning Video Object Segmentation. Object detection and segmentation from videos is closely related to salient object detection. Existing methods rely either on appearance features [29] or motion features from optical flow [6, 35]. One line of research specializes on unsupervised motion segmentation [39], mostly from optical flow [6, 78, 35, 74, 69]. While this opens the avenue to train on large unlabeled datasets, training from 2D optical flow alone does not resolve degenerate motion cases. Other recent work focuses on leveraging vision transformers for generic object discovery [68, 3, 4, 53, 16, 20, 30, 53]. They focus either on unsupervised motion segmentation, video segmentation or generic object feature learning, where motion segmentation potentially acts as input [4]. Their training objectives are not aligned with the presented task definition of [18, 44], where incomplete motion patterns should result in complete semantic object instances. Therefore, we focus on supervised motion segmentation in this work.

Many older approaches have focused on a binary foreground/background separation [57, 19, 40], which would require additional post-processing in order to detect individual objects. Another line of work utilizes binary motion segmentation as an auxiliary task for monocular scene reconstruction [66, 67, 79, 49, 82, 37]. While this achieves promising results, it showcases the chicken-and-egg nature of the problem: In order to reconstruct video, we would like to separate the scene into dynamic foreground and static background beforehand. On the flipside, we need accurate 3D motion fields to infer this assignment in retrospective.

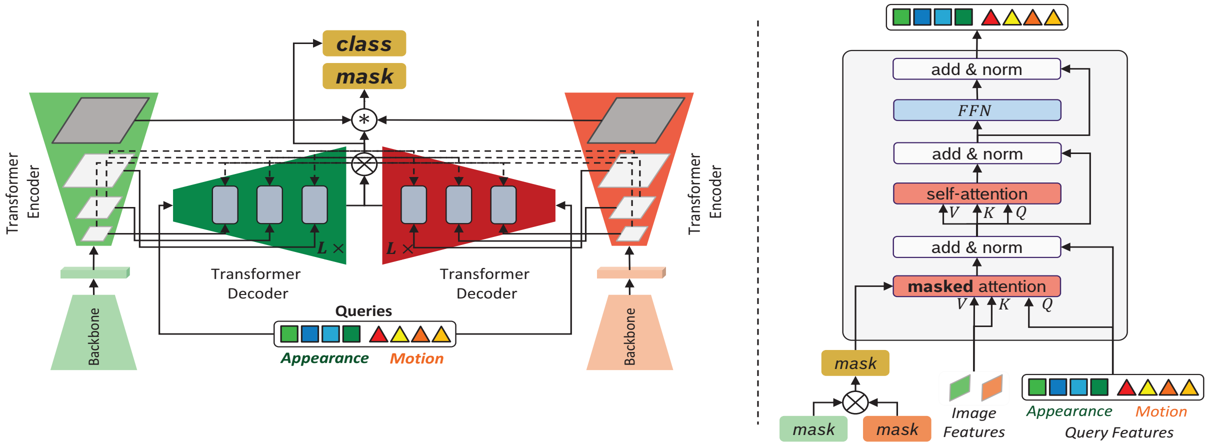

Figure 2: The M^{3}Former architecture. Using two multi-modal streams, we fuse separate image and motion features across the streams. For each stream, we apply a backbone to learn multi-scale features z. We have two separate sets of query embeddings, i.e motion and appearance. Both multi-scale features z and query embeddings q interact with each other through the attention mechanism.

In this work, we use a generic top-down approach to learn instance segmentation in an end-to-end manner similar to [18, 76, 44]. Video Instances Segmentation (VIS) [77] is a highly active research topic for video data [27]. While [18, 68] extend their model to an online-tracker, we focus on instance segmentation in this work. Extending motion segmentation to VIS would be an exciting avenue for future research.

Multi-modal Fusion. Advances in motion segmentation are closely related to instance segmentation. Recent detectors [13, 85, 31] achieve strong performance due to training on large, standard datasets such as COCO [34] and leveraging newer transformer architectures [12, 84, 13]. Pure image based detectors are limited to a fixed number of object labels in the training set. In the same manner, large amounts of data would be needed in order to train a robust motion detector based on image data alone. This can be alleviated by leveraging inductive biases from explicit motion estimates. Similar to [18, 68, 44] we aim to generalize to arbitrary object categories by fusing both appearance and motion information. Motion segmentation is therefore closely related to multi-modal fusion. Since we only have monocular video as input, we create non-rgb *pseudo-modalities* with off-the-shelf expert networks for optical flow [54], depth [47] or scene flow [56]. This approach shares similarities to multi-modal vision expert models [36] or recent multi-modal segmentation transformer [85, 31].

Prior work focused on fusion with CNN-architectures [57, 19, 68, 18, 44]. Fusing with convolutional layers has the downside that both modalities/features will influence each other in a fixed manner. This inflexibility can worsen performance when the information from one modality is corrupted. Furthermore, switching between modalities or extending the architecture from image data to video data cannot be done in a CNN without retraining. We adress these issues by using a transformer with a two-stream architecture consisting of an appearance and motion branch. Similar to [42, 83], we fuse features flexibly based on attention [63]. However, instead of using a shared decoder we fuse features at multiple locations in the network. Compared to prior work, we further fuse multi-scale features in order to achieve higher-resolution masks instead of single-scale features. Finally, our work is also closely related to [43], in the sense that we analyze the effect of different fusion mechanisms on downstream task performance. However, instead of focusing on audio-visual classification, we perform motion segmentation.

3. Our Approach

We introduce the M^{3}Former architecture for this task as is illustrated in Figure 2. The main idea of our approach is to flexibly fuse multi-scale features from both appearance and motion data with attention.

3.1. Motion Representations

While previous work has explored the use of optical flow [18, 74, 69] and higher level rigid motion costs [76, 44], a detailed comparison of different motion representations for a single architecture has not been conducted. We progressively explore segmentation performance depending on the motion representation as input data. We analyze both the performance of single-modality inference and fusion with appearance features. Given two images $I_1,\ I_2 \in \mathbb{R}^{H\times W\times 3}$, we are interested in the motion $F_{1\mapsto 2}$ between both frames.

Optical Flow. Optical flow is a 2D translation field $F \in \mathbb{R}^{H\times W\times 2}$. We use RAFT [54] in our work and take a robust version provided by [44].

Higher-dimensional Motion Costs. Optical flow is a 2D projection of the actual 3D motion. Multiple motions can map to the same projection, therefore the reconstruction is ambiguous. Reconstructing object and camera motion from optical flow has multiple degenerate cases [76]. Degenerate cases appear commonly in applications, e.g. all vehicles on a road drive colinear. In order to detect moving objects robustly, we need some form of 3D prior independent from Structure-from-Motion. The authors of [76] formulate four handcrafted criteria for computing a higher dimensinal cost function $C_{12} \in \mathbb{R}^{H \times W \times 14}$ between two frames. This cost function has a higher cost in regions, that violate the static scene assumption. Computation involves estimating optical flow [54], optical expansion [75], camera motion [24] and monocular depth [47]. The authors of [44] extend this cost function to a three-frame formulation $C_{13} \in \mathbb{R}^{H \times W \times 28}$ by using backward $F_{2 \mapsto 1}$ and forward motion $F_{2 \mapsto 3}$. The computation of this cost embedding involves up to four neural networks, each trained on their own specific datasets.

Scene Flow. There exists a simpler minimal formulation - 3D scene flow. Given two RGBD frames $\{I_1,\ Z_1\}$ and $\{I_2,\ Z_2\}$, we compute motion as a field of rigid body transformations $F \in \mathbb{R}^{H \times W \times 6} \in SE_3$. RAFT-3D [56] is the direct 3D equivalent of the 2D optical flow network [54] and naturally includes a geometric optimization. The main idea of this work is to compute a motion $g \in SE_3$ for each pixel without making any assumption about semantics. Pixels naturally group together into semantically meaningful objects due to moving with the same rigid body motion. We spin this idea around - given multiple rigid body motions in a scene we want to infer an instance segmentation. While there are many diverse datasets for optical flow training [1, 48, 10, 22], there are fewer datasets for scene flow training [41]. We found, that existing model weights do not transfer well to all of our training datasets. We therefore finetune RAFT-3D for our training data, but use published checkpoints [56] during the evaluation. Performance of 3D motion estimation is largely dependent on the depth map quality. Training is done mostly with high-quality or ground truth depth. During inference on in-the-wild data, we do not have access to accurate absolute scale monocular depth for both $Z_1,\ Z_2$. We ablate the performance of motion estimation and segmentation depending on the depth quality.

3.2. Fusion

Image based detectors can solve the segmentation and detection task well, but perform poorly on motion classification. Simply using monocular video data for motion segmentation is a challenging task to learn with limited training data. The task gets solvable when using motion as an intermediate data representation, which acts as inductive biases. However, in order to robustly segment semantically meaningful moving objects, combining both image and motion data together is crucial. The motion segmentation task therefore can be considered a *multi-modal fusion* problem.

Transformers are very flexible - Adapting a transformer for example to Video Instance Segmentation only requires a change in Positional Encoding and little finetuning [13]. This flexibility is a key advantage, since it leaves the possibility open to use longer temporal windows in the future. In a similar manner, we add a modality specific positional encoding and combine data from multiple modalities instead of temporal frames. When using multiple modalities, we combine features within a two-stream architecture with dedicated parameters $\Theta_{rgb},\ \Theta_{motion}$. Each branch is trained on it's own modality individually first and then fusion is learned by finetuning both branches together. We experiment with multiple methods for fusing information at different locations. We base our different streams on the SotA segmentation architecture Mask2Former [14].

Multi-headed Attention. A transformer layer consists of Multi-Headed Self-Attention (MSA) [63], Layer Normalisation (LN) and Multilayer Perceptron (MLP) blocks, applied using residual connections. Given input tokens z^l at layer l, we have

$$y^l = MSA\left(LN\left(z^l\right)\right) + z^l \tag{1}$$

$$z^{l+1} = MLP\left(LN\left(y^l\right)\right) + y^l \quad . \tag{2}$$

The MSA operation computes dot-product attention [63], where query, key and values are linear projections of the same input tensor: $MSA\left(\mathbf{X}\right) = Attention\left(\mathbf{W}^Q\mathbf{X},\ \mathbf{W}^K\mathbf{X},\ \mathbf{W}^V\mathbf{X}\right)$. Multi-Headed Cross Attention (MCA) computes attention between two input tensors $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ acts as the query and $\mathbf{Y}$ as keys and values: $MCA\left(\mathbf{X},\ \mathbf{Y}\right) = Attention\left(\mathbf{W}^Q\mathbf{X},\ \mathbf{W}^K\mathbf{Y},\ \mathbf{W}^V\mathbf{Y}\right)$. Fusion in a vision transformer architecture is simple: Given two separate token sequences z_{rgb} and z_{motion}, we can generate a longer sequence $z = [z_{rgb}||z_{motion}]$ by concatenation. Running this longer sequence through the transformer layer lets both modalities exchange information. We have both *self-attention* and *cross-attention* layers with a learned attention mask $\mathbf{M}^{l-1}$ [15] in the decoder. Since it is a query based detector, we not only have high-resolution spatial input feature tokens z (see Figure 2), but also 256-dimensional object query embeddings q. Masked cross-attention is computed between z and q, while self-attention is performed only on q to learn global context. We have two sets of object embeddings: *appearance* q_{rgb} and *motion* q_{motion}. We concatenate spatial features $[z_{rgb}||z_{motion}]$, object query embeddings $[q_{rgb}||q_{motion}]$ and the respective attention masks $[\mathbf{M}_{rgb}||\mathbf{M}_{motion}]$ as can be seen in Figure 2 on the right. Attention can flow freely through the network with the learned masks, i.e. all

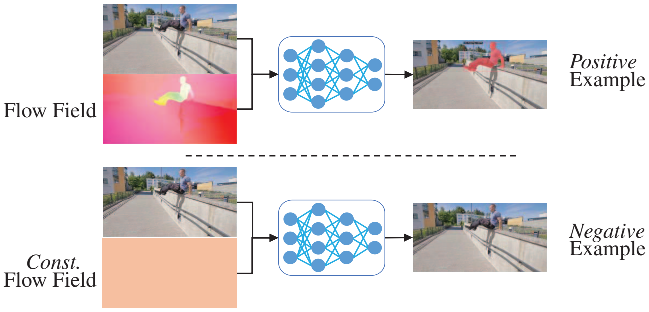

Figure 3: Alignment from regularization: We force the model to not overrely on appearance data by introducing neg. examples.

important image tokens can interact with motion tokens and all queries. A final prediction is made by combining individual outputs with a single convolutional layer.

Attention Bottlenecks. Pairwise attention has quadratic complexity, which can be critical. In order to tame this complexity, the authors of [43] proposed fusion bottleneck tokens. We found that in practice memory only becomes a problem when using many queries. In the same manner, we experimented with bottleneck object embeddings q_{mbt} and let both branches interact only through these bottlenecks.

Deformable Attention. The encoder of the Mask2-Former architecture uses multi-scale deformable attention [84]. This is a mechanism for sampling only few interesting spatial locations from input features maps. The pairwise attention is thus limited to a reduced set and has linear complexity w.r.t the spatial input size. For fusion, we can simply concatenate both feature maps along the x-axis, such that $f = [f_{rgb}, f_{motion}]$ since both input modalities share the same spatial dimensions. When fusing at the encoder level, we perform this operation on each scale of the feature pyramid and add a modality specific positional encoding similar to [13].

Multi-modal Alignment. Alignment between modalities is a vast and important topic in multi-modal models, we refer the reader to [2] for a more extensive overview. Both modalities might not contain the same amount of information and models need to be able to flexibly decide which data source to trust. In our case motion maps might be noisy and the model needs to figure out to rely on the appearance information for high segmentation quality. At the same time datasets [33] exist, where motion can act as a stronger cue for discovering moving objects.

	FT3D	Monkaa	Driving	Vkitti	Kitti	Davis	YTVOS
Mix 0	✓	✗	✗	✗	✗	✗	✗
Mix 1 [76, 44]	✓	✓	✓	✗	✗	✗	✗
Mix 2	✓	✗	✗	✓	✗	✓	✗
Mix 3	✓	✓	✓	✓	✓	✓	✗
Mix 4 [18]	✓	✗	✗	✗	✗	✓	✓

Table 2: We experiment with different dataset mixes. Colored datasets are used for evaluation. Single-modality models are trained on Mix 0, Fusion models on Mix 1 -3. Mix 1 is a common setting proposed by [76, 44] to test generalization. Because this setting lacks diversity in semantic classes and motion patterns, we propose more appropriate mixes that resolve common failure cases.

We notice in our experiments in Section 4, that models usually overrely on appearance data for motion segmentation and thus introduce many false positives. This issue is especially present for ill-posed 2D motion representations. We thus propose a very simple augmentation strategy as can be seen in Figure 3: With a given probability p_{neg}, we introduce negative examples, where motion data is augmented to a random constant flow field within value range. Without any variation in the motion data, models should place semantic objects from the appearance stream into the background. We experiment with multiple values for p_{neg}.

4. Experiments

In our experiments we want to answer the following research questions:

> What motion representation is most useful for motion segmentation? How important is fusion with appearance data?

We use a vanilla Mask2Former [14] model for single-modality training. All experiments are done with a ResNet50 [26] backbone, so that we are comparable to related approaches. Scaling the network is not focus of this paper, but would be a promising direction for future work.

Dataset	Groundtruth data	Diversity Motion	Diversity Classes	non-rigid motion	degenerate cases	#Train	#Test
FlyingTings3D	Depth, 2D/3D Motion, Odometry	High	High	✗	✗	40 100	7800
Monkaa	Depth, 2D/3D Motion, Odometry	Medium	Medium	(✓)	✗	23 356	2588
Driving	Depth, 2D/3D Motion, Odometry	Low	Low	✗	✓	9954	1106
Davis	-	Medium	High	✓	✗	2232	1620
Kitti	Lidar, 2D/3D Motion, Odometry	Low	Low	✗	✓	180	20
Virtual Kitti	Depth, 2D/3D Motion, Odometry	Low	Low	✗	✓	29 811	3314

Table 1: Motion segmentation datasets. Available datasets have different motion patterns and moving semantic classes.

4.1. Datasets

The authors of [18, 76, 44] have made the effort to create motion labels on multiple datasets. Table 1 shows used motion segmentation datasets and their characteristics. We use common datasets: Sceneflow [41], KITTI [22], Virtual Kitti [11] and Davis [46]. Scenes range from autonomous driving, random synthetic scenes to real world casual videos with humans and animals. Table 2 shows different training data mixes from the literature and our experiments. Related work [76, 44] train their fusion models solely on the SceneFlow datasets and evaluate generalization on Davis, Kitti and YTVOS [71]. We drop YTVOS, because performance heavily correlates to Davis. We keep this training setting for our fusion experiments. Single-modality motion segmentors are trained on FlyingThings3D. We note how common failure cases result due to a lack of diverse training data. Mix 1 does not contain many degenerate motion patterns and non-rigid moving objects. We therefore progressively diverge from this setting and analyze the effect of data on performance in Section 4.4. We balance individual datasets, such that samples are drawn with approx. equal likelihood during training, i.e. we use a naive sampling strategy. We believe this to be a step in the right direction, as large scale training is necessary for true real-world generalization abilities.

Metrics. We report standard instance segmentation COCO metrics such as mAP, AP_{50}. We further include other segmentation metrics, such as Precision (Pu), Recall (Ru) and F-score (Fu) [18], foreground precision [76] and the number of false positives and false negatives over the whole split [44]. Since datasets come in different sizes, we normalize the number of false positives/negatives. In our ablations, we mainly focus on mAP, FP and FN, because they act as a good proxy. More details can be found in Suppl. Sec. 7.2.

4.2. Modalities for Motion Segmentation

In our first experiments, we focus on single modalities. We train for 30 epochs, for more details see Suppl. Sec. 7.1. Table 3 shows the results on the test split of FlyingThings3D. We achieve best results with 3D input data, which suggests that 3D motion makes the task easier for the network to learn and generally outperforms 2D motion. The gap between predicted and groundtruth motion leaves room for improvement for off-the-shelf estimators. Interestingly, we include a pure image baseline model. We can train a strong image detector on this dataset, because foreground objects are consistently in motion and distinct from the background. Note how this would not be the case if the data contained object classes, which can move but don't. We will later see, how pure image baselines only perform favorably on metrics which do not punish false positives.

Modality	AP	AP_{50}	AP_{75}
RGB	56.53	76.71	57.5
Scene Flow†	**75.19**	**89.52**	**77.03**
Optical Flow†	72.24	87.43	74.52
Scene Flow [56]	55.39	75.31	56.26
Motion embedding [44]	53.30	75.20	54.9
Optical Flow [54]	52.45	72.75	52.73

Table 3: Comparison of different input data for motion segmentation on FlyingThings3D. † denotes ground truth data.

4.3. Why One Modality Is Not Enough

When generalizing to real-world data with a very different distribution of objects and motion patterns, single-modality models will perform much worse as can be seen in Table 4. For our pure image baseline, we use the COCO [34] pretrained model from [14]. In order to create a stronger baseline, we only use classes, which can move on their own or are likely to be in motion, e.g. cars or persons (see more information in Suppl. Sec. 7.2). 3D motion requires 3D geometry. Monocular depth prediction in dynamic environments is an open problem [32, 81] and is challenging on in-the-wild data. During training we used perfect ground truth depths for computing the scene flow. On in-the-wild data this will not be the case. We ablate multiple scenarios for depth prediction quality. For autonomous driving data we compare the performance for rel. monocular depth, abs. monocular depth and stereo depth. For monocular depth prediction we take DPT [47] and UniMatch [70] for stereo as two SotA single-timeframe models. We compute the abs. depth of each frame by aligning it with the groundtruth as [76]. Alignment is not possible on casual video clips like DAVIS without a reference. The reconstruction of casual videos is still an open research problem in itself [37]. However, we propose a simple strategy for depth alignment based on an end-to-end SLAM system

	Kitti			Davis		
Modality	AP_{50} ↑	FP ↓	FN ↓	AP_{50} ↑	FP ↓	FN ↓
RGB (Coco)	**58.2**	1.34	**0.17**	**50.51**	0.92	**0.07**
Optical Flow [54]	25.1	0.99	0.43	30.2	0.63	0.13
Scene Flow [56] rel. scale	29.6	0.54	0.42	11.0	**0.24**	0.22
Scene Flow [56] abs. scale	36.8	0.50	0.40	39.84	0.41	0.14
Scene Flow [56] stereo	44.4	**0.10**	0.40	-	-	-
Motion embedding [44]	28.9	0.59	0.43	33.9	0.57	0.13

Table 4: Zero-shot performance of single-modality models on KITTI and Davis. Results in grey are only for few selected videos, where a reconstruction with SfM is possible.

	Kitti			Davis		
p_{neg}	AP_{50} ↑	FP ↓	FN ↓	AP_{50} ↑	FP ↓	FN ↓
0	15.55	**14.42**	107.8	19.846	897.90	263.85
30	**37.16**	23.38	**83.19**	**23.98**	**616.53**	**258.81**

Table 5: Ablation of neg. examples augmentation (FP and FN are not normalized). Experiments were run on Mix 1 with image and optical flow data.

			Kitti			Davis		
Data	Modality	Fusion mechanism	AP_{50} ↑	FP ↓	FN ↓	AP_{50} ↑	FP ↓	FN ↓
Mix 1	RGB + OF	D	37.16	0.12	0.41	23.98	0.39	0.16
		E+D	39.65	0.15	0.40	35.88	0.29	0.15
	RGB + SF	D	27.5	0.05	0.50	19.25	0.30	0.19
		E+D	26.5	0.10	0.49	15.4	0.11	0.23
	RGB + SF*	D	27.6	**0.05**	0.50	21.8	0.38	0.14
		E+D	37.8	0.33	0.38	38.4	0.13	0.17
	RGB + SF**	D	27.6	**0.05**	0.50	-	-	-
		E+D	51.0	0.10	0.35	-	-	-
	RGB + Cost	D	29.51	0.06	0.47	27.47	0.30	0.17
		E+D	47.9	0.22	0.35	33.27	0.49	0.14
Mix 3	RGB + OF	D	70.82	0.32	0.16	54.01	0.13	**0.01**
		E+D	60.88	0.29	0.24	61.12	0.11	0.12
	RGB + SF	D	**72.07**	0.31	**0.16**	58.76	0.11	0.13
		E+D	56.3	0.38	0.27	53.9	**0.08**	0.15
	RGB + Cost	D	68.46	0.26	0.21	58.40	0.09	0.13
		E+D	65.12	0.22	0.23	**64.10**	0.11	0.12

Table 6: Fusion of appearance with different modalities. * denotes abs. scale depth ** denotes stereo depth

[55]. This reconstruction is only possible on few selected video clips, but acts as a proof-of-concept. More information can be found in Suppl. Sec. 7.4.

Motion Data Is Not Equally Useful. It can be seen in Table 4, that motion representations can provide different value depending on the dataset. While optical flow is a generic motion representation, which can be inferred reliably on most datasets, 3D scene flow is heavily dependent on the depth quality. The motion embeddings from [44] offer a great trade-off since they do not require multiple scale-correct depth maps, but still contain 3D costs. Once depth is reliably provided, high quality scene flow gives the best results as can be seen on Kitti. However, there remains a large gap in mAP to the image baseline. Reasons for this gap are multiplefold: i) Davis contains many non-rigid motion patterns. Since this has not been in the training data, the model did not learn to group motions and oversegments the scene. ii) 3D geometry cannot be reliably reconstructed, therefore 3D motion is very noisy. iii) Kitti has fast camera motion and many objects move colinear to it. At the same time most scenes contain both static and moving objects of the same class, which is in contrast to training data. Thus, Optical Flow based detectors have many false positives. 3D motion based detectors are dependent on the depth quality. iv) Often multiple objects share the same forward motion, therefore they are grouped together and the scene is undersegmented. These cases are not present in dataset mix 0 and 1. On the other hand, a pure image detector will detect any semantic object and introduce many false positives.

4.4. Fusion Between Appearance and Motion Data

In order to create robust motion segmentation, we resolve the before mentioned problems by fusing appearance and motion information. Since we want to *retain* semantic object knowledge of an image detector, we freeze the image branch that is pretrained on COCO similar to previous work [18, 44]. We take the pretrained motion branches on FT3D and finetune a fusion model on the respective data mix. Training and implementation details can be found in Supp. Sec 7.1. Alignment between appearance and motion features is very important. The model should not rely too much on appearance to overrule the classification from motion. In Table 5 we show the effect of introducing neg. examples. As can be seen, this simple augmentation can stop the model to rely too much on appearance data and reduces false positives (we show the total number of false positives/negatives over the whole split). On Kitti the number of false positives is harder to reduce, because both positive and negative examples of moving objects are hidden inside a flow field with large variance due to the fast driving motion. We keep 30% neg. examples as augmentation in future experiments. When scaling up to larger dataset mix 3, we set $p_{neg} = 5\%$ in order to reduce training time as a trade-off.

We can choose multiple fusion strategies in our two-stream architecture: i) deformable Attention in Encoder (E). ii) Vanilla attention in Decoder (D). iii) Multi-modal Bottleneck Tokens (MBT) [43] in Decoder. iv) Fusion in both Encoder and Decoder (E+D). We ablate these strategies in Table 11 in Suppl. Sec. 8.1. We found, that there is no optimal strategy for all motion representations and training data. We observed, that the training dynamics are affected by the fusion mechanism and hypothesize, that the strategies can potentially converge to similar results when given enough training time. Finally, there is no optimal strategy for both Kitti and Davis. We therefore opted for the simple late fusion in the decoder or fusion in both encoder and decoder for later experiments. Our results in Table 6 show, that motion cues generally reduce false positives and the fusion with appearance data closes the gap in precision. 3D motion representations can give stronger performance when they are available in high quality.

Beyond Small-scale Datasets. Our previous experiments have shown that a simple detector baseline is hard to beat for segmentation precision. While image data is very valuable for precision, motion data helps in reducing false positive detections as can be seen in Table 4. Motion and Fusion models over- or undersegment the scene due to a lack of diverse training data. As can be seen in Table 7 we could not replicate the performance of [44] with the training setting of Mix 1 [44, 76]. Mix 1 does neither contain real

		Kitti								Davis							
Method	Training data	*AP* ↑	bg	obj	Pu	Ru	Fu	FP ↓	FN ↓	*AP* ↑	bg	obj	Pu	Ru	Fu	FP ↓	FN ↓
RGB sem. baseline [13]	COCO	42.2	96.6	69.25	60.70	93.96	69.25	1.34	0.17	35.05	0.92	0.68	0.61	**0.88**	0.68	0.92	0.07
Learning rigid motions [76] *	Mix 1	20.0	-	-	-	-	-	-	-	4.2	-	-	-	-	-	-	-
Generic MoSeg[18]	Mix 4	20.0	-	-	-	-	-	-	-	20.8	-	-	-	-	-	-	-
Raptor [44]	Mix 1	40.07	98.97	86.3	**89.37**	86.3	86.3	**0.11**	0.35	40.9	94.20	73.3	71.57	80.20	73.3	0.25	0.10
Ours RGB + OF	Mix 1	25.67	98.46	76.63	81.12	76.70	76.63	0.15	0.40	19.28	**95.34**	64.85	67.32	67.82	64.85	0.29	0.15
Ours RGB + SF*		26.70	96.65	64.13	62.71	77.08	64.13	0.33	0.38	15.20	96.50	69.45	71.07	69.70	69.45	0.13	0.17
Ours RGB + Cost [44]		32.40	98.65	79.39	78.47	84.87	79.39	0.22	0.35	16.85	93.70	59.78	62.39	64.67	59.78	0.50	0.14
Ours RGB + OF	Mix 2	40.08	97.83	66.84	74.59	68.10	66.84	0.13	0.35	**43.52**	92.97	**76.72**	76.62	83.48	**76.72**	0.25	0.09
Ours RGB + OF	Mix 3	50.91	**99.19**	85.99	83.26	91.30	85.99	0.32	0.16	32.25	94.53	73.73	73.12	77.23	73.73	0.12	**0.01**
Ours RGB + SF	Mix 3	**52.27**	98.89	**87.05**	84.18	**93.99**	**87.05**	0.31	**0.16**	37.07	95.17	76.21	**77.24**	77.70	76.21	0.11	0.13
Ours RGB + Cost [44]	Mix 3	48.44	98.76	82.33	80.84	88.13	82.33	0.26	0.21	35.11	95.13	75.87	75.58	78.51	75.87	**0.09**	0.13

Table 7: SotA Motion Segmentation on Kitti and Davis. We report our best results for the respective modality and data. Results in grey are on scenes, where a reconstruction with SfM is possible. *use of abs. scale information

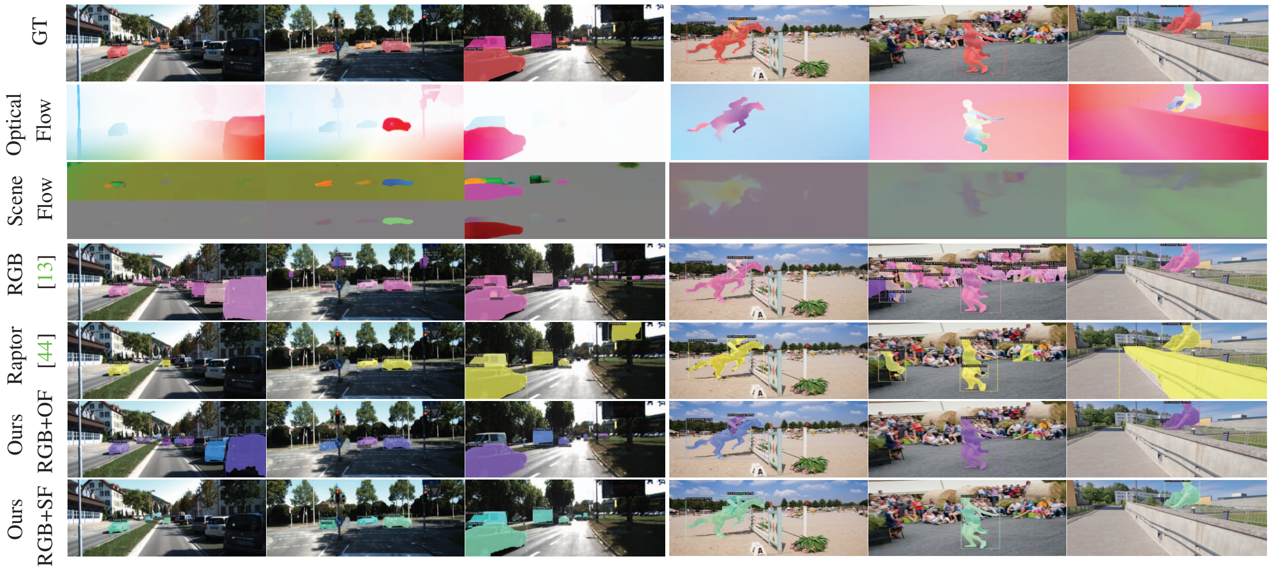

Figure 4: Qualitative comparison on Kitti and Davis.

data with non-rigid motions nor realistic driving scenes with many traffic participants. Since a variety of datasets exists, we can fix this problem by training on more diverse data. We combine sources from up to six datasets in our training. Our new mixes 2 and 3 offer multiple cases of non-rigid motions, multiple objects moving in union and hard degenerate motion scenarios. See Figure 7 in Appendix for examples of how different training data can improve previous failures. It can be seen in Table 11 and 6, how both training data and different modalities affect performance. We observe that depending on the training data, performance of $\mathbf{M}^3$Former improves drastically (see Figure 8 in Suppl. Sec. 9.2).

Table 7 shows the SotA in supervised motion segmentation on Kitti and Davis. We visualize examples of the test splits in Figure 4. Note how for Mix 3 results, the models still have never seen the evaluation data. Our results are not necessarily surprising, as we partially trained on the target domain. However, we find that our incremental improvements behave quite causal: *Most failure modes of the model disappear when supervised properly*. Previous shortcomings can be resolved solely with better datasets instead of architectural changes. Our proposed model architecture is simple and flexible. Surprisingly, our results show that even by using 2D optical flow, we can reach SotA performance on Kitti without using any real driving data. It can be seen that a minimal 3D motion representation like scene flow can be effective even with noisy data. Models can pick up strong cues for moving objects even from context alone. For example, a car that is placed on a driving lane is likely in motion compared to one parked to the side. Interestingly, creating a balanced dataset is a new optimization problem in itself [47]. We observe, that adding just more data sources can detoriate performance on Davis. Depending on the downstream-application, the data needs to be correctly balanced. We leave this for future work.

5. Conclusion

We systematically analyzed the motion segmentation problem from monocular video. In our experiments we identified the importance of different 2D and 3D motion representations on multiple datasets. We proposed a novel transformer fusion architecture $\mathbf{M}^3$Former which fuses appearance and motion information on multiple scales. We analyzed multiple fusion schemes within this framework. Our approach achieves SotA performance by leveraging the flexible attention mechanism and diverse training data. Our findings showed that both 2D and 3D motion can give strong performance when trained on appropriate data. Since appearance data mostly drives segmentation, the importance of high-quality motion estimates gets weaker when scaling the data size.

ACKNOWLEDGMENT. Research presented here has been supported by the Robert Bosch GmbH. We thank our colleagues Annika Hagemann, Fabian Schmidt, Tamas Kapelner and EashwaraErahan for fruitful discussions.

References

[1] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92:1–31, 2011. 4

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 5

[3] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering objects that can move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11789–11798, 2022. 2, 13

[4] Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Object discovery from motion-guided tokens. *arXiv preprint arXiv:2303.15555*, 2023. 2, 13

[5] 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: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 433–449. Springer, 2016. 2, 13

[6] Pia Bideau, Rakesh R Menon, and Erik Learned-Miller. Moa-net: self-supervised motion segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 13

[7] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 508–517, 2018. 13

[8] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4322–4331, 2019. 2

[9] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V 11*, pages 282–295. Springer, 2010. 2

[10] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 4

[11] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 6

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3, 13

[13] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1, 3, 4, 5, 8, 19

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 4, 5, 6, 13, 14

[15] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 4

[16] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess what moves: unsupervised video and image segmentation by anticipating motion. *arXiv preprint arXiv:2205.07844*, 2022. 2, 13

[17] Trevor Darrell and Alexander Pentland. Robust estimation of a multi-layered motion representation. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 173–174. IEEE Computer Society, 1991. 2

[18] 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. 1, 2, 3, 5, 6, 7, 8, 13, 14

[19] 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. 2, 3

[20] Gamaleldin Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C Mozer, and Thomas Kipf. Savi++: Towards end-to-end object-centric learning from real-world videos. *Advances in Neural Information Processing Systems*, 35:28940–28954, 2022. 2, 13

[21] Katerina Fragkiadaki and Jianbo Shi. Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement. In *CVPR 2011*, pages 2073–2080. IEEE, 2011. 2

[22] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 4, 6

[23] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992. 1

[24] Richard I Hartley. In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 19(6):580–593, 1997. 4

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 14

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 13

[27] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *arXiv preprint arXiv:2206.04403*, 2022. 3

[28] Michal Irani, Benny Rousso, and Shmuel Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Computer Vision—ECCV'92: Second European Conference on Computer Vision Santa Margherita Ligure, Italy, May 19–22, 1992 Proceedings 2*, pages 282–287. Springer, 1992. 2

[29] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Pixel objectness. *arXiv preprint arXiv:1701.05349*, 2017. 2

[30] Laurynas Karazija, Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised multi-object segmentation by predicting probable motion patterns. *arXiv preprint arXiv:2210.12148*, 2022. 2, 13

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[32] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 6

[33] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 5, 16, 19

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 6, 14

[35] Runtao Liu, Zhirong Wu, Stella Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. *Advances in Neural Information Processing Systems*, 34:13137–13152, 2021. 2, 13

[36] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*, 2023. 3

[37] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. *arXiv preprint arXiv:2301.02239*, 2023. 1, 2, 6, 15

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 13

[39] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3623–3632, 2019. 2, 13

[40] Zhaoyang Lv, Kihwan Kim, Alejandro Troccoli, Deqing Sun, James M Rehg, and Jan Kautz. Learning rigidity in dynamic scenes with a moving camera for 3d motion field estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 468–484, 2018. 2, 13

[41] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 4, 6, 14

[42] Eslam Mohamed and Ahmad El-Sallab. Modetr: Moving object detection with transformers. *arXiv preprint arXiv:2106.11422*, 2021. 3, 13

[43] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 3, 5, 7, 14, 15, 16

[44] Michal Neoral, Jan Šochman, and Jirí Matas. Monocular arbitrary moving object discovery and segmentation. 2021. 1, 2, 3, 4, 5, 6, 7, 8, 13, 14, 18, 19

[45] 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, 2013. 2

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

[47] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1, 2, 3, 4, 6, 8, 15, 17

[48] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 4

[49] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 2

[50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 13

[51] Harpreet S Sawhney. 3d geometry from planar parallax. In *CVPR*, volume 94, pages 929–934, 1994. 2

[52] Jianbo Shi and Jitendra Malik. Motion segmentation and tracking using normalized cuts. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 1154–1160. IEEE, 1998. 2

[53] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *arXiv preprint arXiv:2205.14065*, 2022. 2, 13

[54] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1, 2, 3, 4, 6, 14

[55] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 7, 15

[56] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8375–8384, 2021. 1, 2, 3, 4, 6

[57] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4490, 2017. 2, 3, 13

[58] Philip HS Torr. Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 356(1740):1321–1340, 1998. 2

[59] Philip HS Torr, Andrew W Fitzgibbon, and Andrew Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. *International Journal of Computer Vision*, 32:27–44, 1999. 2

[60] Philip HS Torr, Andrew Zisserman, and Stephen J Maybank. Robust detection of degenerate configurations while estimating the fundamental matrix. *Computer vision and image understanding*, 71(3):312–333, 1998. 2

[61] Roberto Tron and René 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. IEEE, 2007. 2

[62] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016. 2

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[64] René Vidal and Richard Hartley. Motion segmentation with missing data using powerfactorization and gpca. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004. 2

[65] René Vidal and Shankar Sastry. Optimal segmentation of dynamic scenes from two perspective views. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE, 2003. 2

[66] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 2

[67] Jonas Wulff, Laura Sevilla-Lara, and Michael J Black. Optical flow in mostly rigid scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4671–4680, 2017. 2

[68] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9994–10003, 2019. 2, 3, 13

[69] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *Advances in Neural Information Processing Systems*, 2022. 2, 3

[70] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *arXiv preprint arXiv:2211.05783*, 2022. 6

[71] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 6, 17

[72] Xun Xu, Loong-Fah Cheong, and Zhuwen Li. 3d rigid motion segmentation with mixed and unknown number of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):1–16, 2019. 2

[73] Jingyu Yan and Marc Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*, pages 94–106. Springer, 2006. 2

[74] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7188, 2021. 2, 3, 13

[75] Gengshan Yang and Deva Ramanan. Upgrading optical flow to 3d scene flow through optical expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1343, 2020. 2, 4

[76] Gengshan Yang and Deva Ramanan. Learning to segment rigid motions from two frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1266–1275, 2021. 2, 3, 4, 5, 6, 7, 8, 13, 14

[77] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 3

[78] 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. 2

[79] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[80] Chang Yuan, Gerard Medioni, Jinman Kang, and Isaac Cohen. Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1627–1641, 2007. 2

[81] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 6

[82] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 523–542. Springer, 2022. 1, 2

[83] 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*, volume 34, pages 13066–13073, 2020. 3, 13

[84] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3, 5, 13

[85] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 3