Motion and Appearance Based Multi-Task Learning Network for Autonomous Driving

Mennatullah Siam
University of Alberta
mennatul@ualberta.ca

Heba Mahgoub
Cairo University
h.mahgoub@fci-cu.edu.eg

Mohamed Zahran
Valeo Deep Learning Research
mohamed.zahran@valeo.com

Senthil Yogamani
Valeo Vision Systems
senthil.yogamani@valeo.com

Martin Jagersand
University of Alberta
jag@cs.ualberta.ca

Ahmad El-Sallab
Valeo Deep Learning Research
ahmad.el-sallab@valeo.com

Abstract

Autonomous driving has various visual perception tasks such as object detection, motion detection, depth estimation and flow estimation. Multi-task learning (MTL) has been successfully used for jointly estimating some of these tasks. Previous work was focused on utilizing appearance cues. In this paper, we address the gap of incorporating motion cues in a multi-task learning system. We propose a novel two-stream architecture for joint learning of object detection, road segmentation and motion segmentation. We designed three different versions of our network to establish systematic comparison. We show that the joint training of tasks significantly improves accuracy compared to training them independently even with a relatively smaller amount of annotated samples for motion segmentation. To enable joint training, we extended KITTI object detection dataset to include moving/static annotations of the vehicles. An extension of this new dataset named KITTI MOD is made publicly available via the official KITTI benchmark website [1]. Our baseline network outperforms MPNet which is a state of the art for single stream CNN-based motion detection. The proposed two-stream architecture improves the mAP score by 21.5% in KITTI MOD. We also evaluated our algorithm on the non-automotive DAVIS dataset and obtained accuracy close to the state-of-the-art performance. The proposed network runs at 8 fps on a Titan X GPU using a two-stream VGG16 encoder. Demonstration of the work is provided in [2].

1 Introduction

Autonomous driving is a rapidly advancing application area with the progress in deep learning. There are two main paradigms in this area: (1) The mediated perception approach which semantically reasons the scene [6][17] and then determines the driving decision based on it. (2) The behavior reflex approach that learns end to end the driving decision [2] [22]. The behavior reflex methods can benefit from semantic reasoning of the environment. For example, an auxiliary loss on semantic segmentation [22] was used with end to end learning. On the other hand, in mediated perception semantic reasoning is a central task, followed by the control decision separately. Semantic reasoning of the scene includes object detection, motion detection, depth estimation, object tracking and others. Multiple architectures for the joint reasoning of the different tasks were proposed [17][10]. A shared encoder between these tasks were used, but their work utilizes appearance cues only. Previous

[1] http://www.cvlibs.net/datasets/kitti/eval_semantics.php

[2] https://www.youtube.com/watch?v=hwP_oQeULfc

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

multi-task learning systems do not incorporate motion cues. Although tasks such as motion detection, depth and flow estimation will benefit from it.

Motion detection in particular is a challenging problem because of the continuous camera motion along with the motion of independent objects. Moving objects are the most critical in terms of avoiding fatalities and enabling smooth maneuvering and braking of the car. Motion cues can also enable generic object detection as it is not possible to train for all possible object categories beforehand. Classical approaches in motion detection were focused on geometry based approaches [19][14][13][12][21]. However, pure geometry based approaches have many limitations, motion parallax issue is one such example. A recent trend [18][9][4][20][5] for learning motion in videos has emerged. Nonetheless, this trend was focused on pixel-wise motion segmentation. Fragiadaki et. al. suggested a method to segment moving objects [5] that uses a separate proposal generation. However, proposal generation methods are computationally inefficient. Jain et. al. presented a method for appearance and motion fusion in [9]. The work focuses on generic object segmentation. It was not designed for static/moving vehicles classification. Tokmakov et. al. [18] used a one-stream fully convolutional network with optical flow input to estimate the motion type. The approach works with either optical flow only or concatenated image and flow as input. The concatenated input will not benefit from the available pretrained weights, as they were trained on RGB only. Drayer et. al. [4] described a video segmentation work that used tracked detections from R-CNN denoted as tubes. This was followed by a spatio-temporal graph to segment objects. The main issue with this approach is its running time of 8 seconds per frame. Thus, there is a need for an efficient and more accurate solution.

In this paper, we propose a novel method for scene understanding that combines motion and appearance cues. The motivation behind combining motion and appearance in a multi-task learning system is to improve the benefit for relatively fewer data tasks. Tasks like motion segmentation which have relatively smaller datasets improve through the joint training with detection. The contributions of this work are as follows: (1) We present a novel multi-task learning system for autonomous driving that fuses both appearance and motion cues. (2) This system is used to jointly detect vehicles and segment motion. Another extension of the work had the previous two tasks along with road segmentation. (3) We propose a method to generate automatically annotated data for this task from KITTI dataset which we call KITTI MOD. This provides a benchmark for autonomous driving application, unlike synthetic sequences [11].

2 Creation of KITTI MOD Dataset

Simultaneous annotation is required for jointly training the object detection and motion segmentation. Thus we extended the KITTI object detection dataset to include motion segmentation. We implemented a pipeline to automatically generate static/moving classification for objects which is then manually verified. The procedure uses odometry information and annotated 3D bounding boxes for vehicles. The odometry information that includes GPS/IMU readings provides a method to compute the velocity of the moving camera. The 3D bounding boxes of the annotated vehicles are projected to 2D images and tagged with their corresponding 3D centroid. The 2D bounding boxes are associated between consecutive frames using intersection over union. The estimated vehicles velocities are then computed based on the associated 3D centroids. The computed velocity vector per bounding box is compared to the odometry ground-truth to determine the static/moving classification of vehicles. The objects that are then consistently identified on multiple frames as moving are kept. In this dataset, the focus is on vehicles with car, truck, and van object categories.

An overview of the labeling procedure is shown in Figure 1. This is applied on six sequences from KITTI raw data [7] to generate a total of 1750 frames. In addition to these frames, 200 frames from KITTI scene flow are used to provide us with 1950 frames in total. This new dataset is referred to as KITTI MOD throughout the paper. For some statistics on the dataset, the total number of static vehicles is 5997, while the number of moving ones is 2383. An extension of the dataset is publicly available [1] to act as a benchmark on motion detection on KITTI. Although there exists other motion segmentation datasets such as [15][11][13]. However, they are either synthetic[11], relatively small [13] or has limited camera motion [15] unlike what is present in autonomous driving scenes.

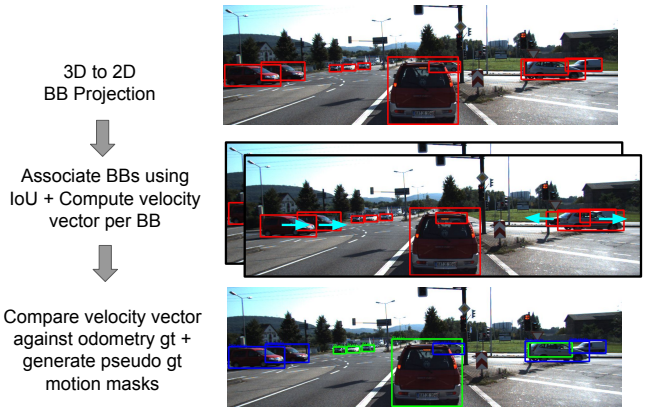

Figure 1: Overview of the pipeline used to generate KITTI Moving Object Detection annotations. Blue boxes for moving vehicles, green boxes for static ones.

3 Two-Stream Multi-task Learning System

In this section both motion and object detection networks are detailed. First a method for jointly detecting vehicles and segmenting motion is presented. Then, a multi-task learning system combining motion and appearance for motion segmentation, vehicle detection and road segmentation is described.

3.1 Joint Vehicle Detection and Motion Segmentation

In autonomous driving, static/moving classification on the object-level is more relevant than dense pixel-level classification. A method that jointly detects vehicles in the scene while classifying them into static/moving is presented. A detector similar to the detection decoder in [17] denoted as FastBox is used. It is based on Yolo [16] used as a single shot detector utilizing the first 15 convolutional layers from VGG16. This is followed by two 1x1 convolutional layers. The last layer outputs 39x12 grid size representing each cell. The channels in the output layer include the bounding box coordinates, size, and the confidence in the existence of a vehicle. The loss function used in detection combines the L1 loss for the bounding box regression, with cross entropy for the confidence score.

A two-stream VGG16 encoder is used to output the combined motion and appearance features. This is followed by two decoders for vehicle detection and motion segmentation. This network is referred to as moving object detection network (MODNet). This method follows a similar approach to the work in [17]. However, in our approach we present motion cues as another valuable input to any multi-task learning network in auto-driving. Our work also has similarities to the work in [9], but their work included one task only for video segmentation. This is one of the main strengths of our work; In the same forward pass motion segmentation and vehicle detection are predicted. This is crucial for real-time performance in autonomous driving scenarios. Inside the segmentation network for each skip connection a summation junction is used to combine motion and appearance features.

$$L_{total} = L_{moseg} + L_{seg} + L_{det} \tag{1a}$$

$$L_{moseg} = -\frac{1}{|I|}\sum_{i\in I}\sum_{c\in C_{motion}} p_i(c)\log q_i(c) \tag{1b}$$

$$L_{seg} = -\frac{1}{|I|}\sum_{i\in I}\sum_{c\in C_{seg}} p_i(c)\log q_i(c) \tag{1c}$$

$$L_{det} = \frac{1}{|S|}\sum_{s\in S} 1^{obj}(|x_{q_s} - x_{p_s}| + |y_{q_s} - y_{p_s}| + |w_{q_s} - w_{p_s}| + |h_{q_s} - h_{p_s}|) \\ -\frac{1}{|S|}\sum_{s\in S}\sum_{c'\in C_{vehicle}} p_s(c')\log q_s(c') \tag{1d}$$

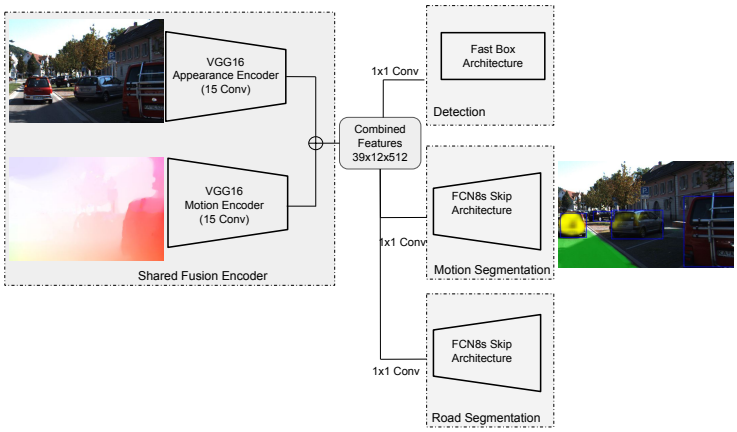

Figure 2: MODNet Two Stream Multi-Task Learning Architecture for joint motion segmentation, object detection and road segmentation. Optical Flow and RGB input, RGB image with overlay: (Yollow) motion segmentation. (Blue) detected bounding boxes. (Green) segmented road.

3.2 Joint Vehicle Detection, Motion and Road Segmentation

Another multi-task learning system is designed that includes three tasks. These are vehicle detection, road segmentation, and motion segmentation. The system provides us with free space information along with alerts to moving detected vehicles. In a similar fashion a shared fusion encoder is used to incorporate both motion and appearance. The output from the shared encoder is used as input to another FCN8s decoder. Thus, in one forward pass vehicle bounding boxes, motion masks and segmented road are predicted. The loss function used in the different multi-task learning system presented alternates between segmentation and detection losses. These losses are shown in Equation 1, where q denotes predictions and p denotes ground-truth. The pixel locations are termed as I, while S is the number of grid cells. C_{motion} is the set of classes for motion segmentation as foreground or background, while $C_{vehicle}$ is the classes for vehicle classification. The detection loss regresses with L1 loss on the coordinates within the cell. Only cells with a positive confidence score are considered in the regression loss. The road segmentation loss is similar to the motion, with two classes. Joint training is performed similarly to [17] where gradients are back-propagated from both tasks on their corresponding mini-batch inputs. This method of joint training leverages the performance of tasks with comparably fewer data. This provides another motivation for the shared motion and appearance encoders. It is worth noting, that motion relevant annotations such as motion masks or optical flow ground-truth are relatively small in real datasets. The tasks for training are selected in an alternate fashion with equal probabilities.

4 Experiments

4.1 Experimental Setup

Throughout experiments, the Adam optimizer is used with learning rate $1e^{-5}$. L2 regularization is used in the loss function to avoid overfitting the data, with $5e^{-4}$ factor. Dropout with probability 0.5 is used to 1x1 convolutional layers. The encoder is initialized with VGG pretrained weights on Imagenet. Transposed convolution layers are initialized to bilinear upsampling. Input image resolution used is 1248x384.

The evaluation metrics used in segmentation are precision, recall, F-score and mean intersection over union (IoU). The evaluation metric used for detection is mean average precision(mAP) and average precision (AP) for static/moving classes. Average precision of car class is also measured showing different difficulties for easy, medium, and hard setup as in KITTI benchmark [8]. Note that it is important to evaluate the static/moving classification standalone without including errors from the detection itself. The average precision used is computed on the detected bounding boxes that match bounding boxes from the ground truth. Thus, evaluation is for static/moving classification standalone, without penalizing errors from FastBox detection.

4.2 Experimental Results

4.2.1 Joint Motion Segmentation and Vehicle Detection

Detailed experiments on motion segmentation with vehicle detection is conducted on KITTI MOD. Table 1 shows the evaluation of the separate and joint training for motion segmentation and vehicle detection. The detection evaluation for the separate setup is taken from [17] since their pre-trained weights are used in this setup. It clearly shows that the joint training improves the motion segmentation with 8.2% approximately in F-score. The detection on the easy evaluation is only affected by 2.5% and on the hard evaluation is approximately the same. It is worth noting that joint training of both tasks improves results when there is limited training data.

Table 1: Quantitative comparison on KITTI MOD data for separate MODNet against jointly trained MODNet.

	Object Detection			Motion Segmentation			
	moderate	easy	hard	Precision	Recall	F-score	IoU
MODNet- Separate	**83.35**	**92.8**	67.59	44.34	69.84	54.25	37.22
MODNet- Joint	80.74	89.52	**67.72**	**56.18**	**70.32**	**62.46**	**45.41**

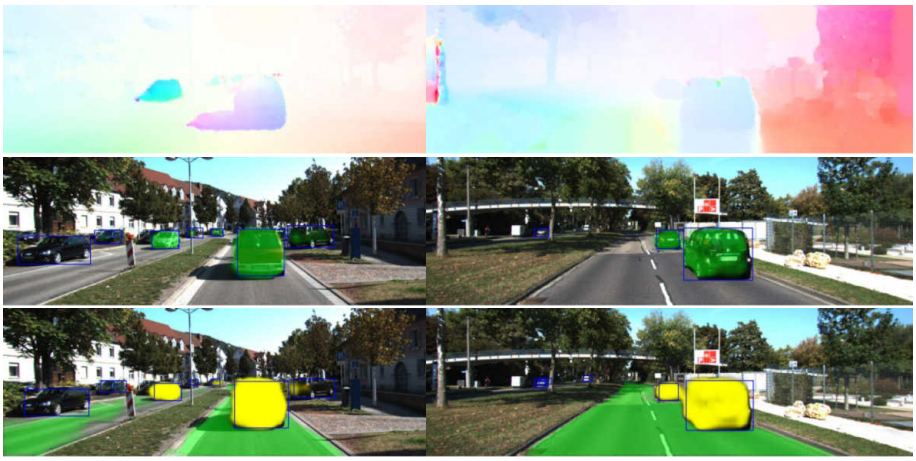

Figure 3: Qualitative evaluation on KITTI MOD data for our proposed two-stream multi-task learning network MODNet. top row: Input Optical Flow, middle row output of 2 tasks: overlay motion mask (green), bottom row output of 3 tasks: overlay motion mask (yellow), road segmentation (green) and detected bounding boxes (blue).

The two-stream motion segmentation network is used to provide motion masks which are then combined with FastBox [17] detections. The output segmentation and vehicles' static/moving classification is evaluated on KITTI MOD data. Table 2 shows the results from the joint detection and motion segmentation. The two-stream MODNet shows the best mAP on KITTI MOD data. This is compared against one of the state-of-the-art methods MPNet [18]. MPNet with optical flow input is evaluated on KITTI MOD and combined with proposals as mentioned in their method. Its pretrained weights are used as is, then its output motion segmentation is used with vehicle detection. If intersection over union is larger than 0.5, the detected vehicle is considered moving. This is applied for both our approach and MPNet. It is worth noting that our method for evaluating static/moving classification does not depend on the object detection itself as explained earlier.

Our proposed approach outperforms MPNet with 21.5% in mAP. This shows that autonomous driving scenarios, exhibit different challenges compared to generic object segmentation. The continuous camera motion and the existence of multiple objects in the scene makes it more challenging. The reasons behind our improvement is two fold. The KITTI MOD training data provide a better representation for motion than the synthetic data used in MPNet. The usage of both optical flow and RGB in a two-stream network that utilizes pretrained VGG16 weights improves the results even more. The two-stream image pair is worse in mAP compared to (RGB+OF), but it is more computationally efficient. The joint detection and motion segmentation method provides an efficient way to solve

both tasks. Our method runs at 8 fps on a TITANX GPU. This outperforms other approaches in the literature in terms of computational efficiency. The running time for approaches that estimate scene flow can be up to 50 minutes, while the approach in [4] takes up to 8 seconds per frame.

Table 2: Quantitative evaluation on KITTI MOD data for our proposed joint detection and motion segmentation network.

	AP Static	AP Moving	mAP
MPNet[18]	50.23	31.84	41.03
MODNet (image pair)	60.7	44.29	52.5
MODNet (RGB+OF)- Separate	**65.28**	56.86	61.07
MODNet (RGB+OF)- Joint	58.6	**66.54**	**62.57**

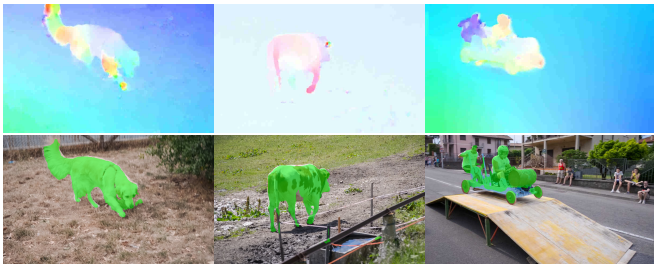

Figure 4: Qualitative evaluation on DAVIS for our proposed two-stream motion segmentation network. RGB Image, Optical Flow and Overlay Motion mask in green.

Table 3: Quantitative evaluation on Davis[15] data Val 2016 using mean IoU. Approaches highlighted in blue are without CRF post-processing, and in red after post-processing.

	MSG[3]	FST[14]	BMM[21]	MPNet[18]	MPNet[18]+CRF	ours	ours+CRF
mIoU	53.3	55.8	62.5	62.66	**70.0**	**63.88**	66.0

4.2.2 Generic Motion Segmentation on DAVIS

To additionally compare against the state of the art in segmentation, our method is evaluated on the Davis[15] benchmark. MODNet is trained on DAVIS training data and evaluated on the validation set. Then it is compared to the unsupervised methods on DAVIS video segmentation benchmark. Note that on DAVIS the term unsupervised denotes that no masks from the initial frame is used as initialization. MPNet is one of the unsupervised methods that works with one stream only and optical flow as input. It is evaluated with and without applying conditional random fields as a post processing, and with the usage of optical flow only. Table 3 shows that our method outperforms the state of the art on DAVIS in unsupervised motion segmentation, except for MPNet+CRF. The improvement over MPNet alone is only 1.5%. MPNet+CRF performs better than ours+CRF, but conditional random field runs in 1.15 seconds per frame. This was measured using input image resolution of 480x854 on an Intel core i5 CPU at 2.30 GHZ. Hence, the usage of CRF as postprocessing is impractical for real-time autonomous driving tasks.

The DAVIS data has very simple camera motion compared to KITTI. Another difference between KITTI sequences and DAVIS is that moving objects cover large portions of the scene. Thus, using optical flow can be sufficient for segmentation. Figure 4 shows the optical flow and segmentation output from our approach on DAVIS data.

5 Conclusion

In this paper, we design a novel multi-task learning system for autonomous driving that combines motion and appearance cues. The system jointly estimates the motion mask, object detection, and road segmentation. Experimental results show that the combined appearance and motion cues in a multi-task learning system outperforms other architectures. Our approach outperforms the single-stream state-of-the-art MPnet by 21.5% in mAP on the extended KITTI dataset (KITTI MOD).

References

[1] KITTI Semantic Segmentation, KITTIMoSeg. `http://www.cvlibs.net/datasets/kitti/eval_semantics.php`, 2017.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[3] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. *Computer Vision–ECCV 2010*, pages 282–295, 2010.

[4] Benjamin Drayer and Thomas Brox. Object detection, tracking, and motion segmentation for object-level video segmentation. *arXiv preprint arXiv:1608.03066*, 2016.

[5] Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving objects in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4083–4090, 2015.

[6] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):1012–1025, 2014.

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[9] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. *arXiv preprint arXiv:1701.05384*, 2017.

[10] Iasonas Kokkinos. Ubernet: Training auniversal'convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *arXiv preprint arXiv:1609.02132*, 2016.

[11] 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.

[12] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[13] 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.

[14] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784, 2013.

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

[16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[17] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv preprint arXiv:1612.07695*, 2016.

[18] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. *arXiv preprint arXiv:1612.07217*, 2016.

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

[20] 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.

[21] Scott Wehrwein and Richard Szeliski. Video segmentation with background motion models. In *BMVC*, 2017.

[22] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. *arXiv preprint arXiv:1612.01079*, 2016.