Real-time Semantic and Class-agnostic Instance Segmentation in Autonomous Driving

Eslam Mohamed[*1], **Mahmoud Ewaisha**[*1], **Mennatullah Siam**[2], **Hazem Rashed**[1], **Senthil Yogamani**[3], **Waleed Hamdy**[1], **Muhammad Helmi**[4] and **Ahmad El-Sallab**[1]
[*]Equal contribution [1]Valeo Egypt [2]University of Alberta [3]Valeo Ireland
[4]Zewail City of Science and Technology

Abstract

Towards a safety critical approach it is of significant importance to perform class agnostic along with semantic instance segmentation. In class agnostic instance segmentation motion cues act as a strong signal to indicate an obstacle regardless of whether it is in the closed set of known classes or not. In this paper, we propose a multi-task model that learns separate heads for semantic and class agnostic instance segmentation to account for both known and unknown objects. Our multi-task model design is computationally efficient through sharing a prototype generation network while learning separate coefficients for each task. In addition, we provide a new public dataset, KITTI-Instance-MoSeg, which increases the number of object categories instead of solely focusing on car to learn a generalizable class agnostic segmentation. We obtain $\sim$ 39 fps with 10% mAP improvement relative to the baseline, while outperforming state of the art methods with 3.3%. We summarize our work in a short video with qualitative results at `https://sites.google.com/view/instancemotseg/`.

1 Introduction

Most of the modern autonomous driving(AD) systems leverage HD maps which perceive the static infrastructure [7, 11], but moving objects are more critical to be detected accurately. Thus motion segmentation is an important perception task for AD [15, 13, 8, 10]. Class agnostic instance segmentation can be seen as an alternate way to semantic instance segmentation and thus providing redundancy needed for a safe and robust system. Depending on motion cues regardless of semantics would scale better to unknown objects since it is practically infeasible to collect data for every possible object category.

In this paper, we build upon Yolact [1] that formulates instance segmentation as learning prototypes which forms a basis of a vector space and then learns the linear combination weights of these prototypes per instance. However, for the purpose of learning both tasks we propose to jointly learn semantic and class agnostic (motion) instance segmentation using a shared backbone and prototype generation (protonet) weights with shallow prediction heads for each task. The semantic and class agnostic heads predict bounding boxes and prototype coefficients. The separation of the two tasks allows for training the class agnostic head on separate datasets that provide class agnostic annotations such as DAVIS [2].

To summarize, the contributions of this work include: (1) The first attempt to perform joint semantic and class agnostic instance segmentation, to the best of our knowledge, that would create a step towards a fail safe systems. Our model maintains real-time performance through sharing the backbone and protonet then learning prototype coefficients per task. (2) We release KITTI-Instance-MoSeg

Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

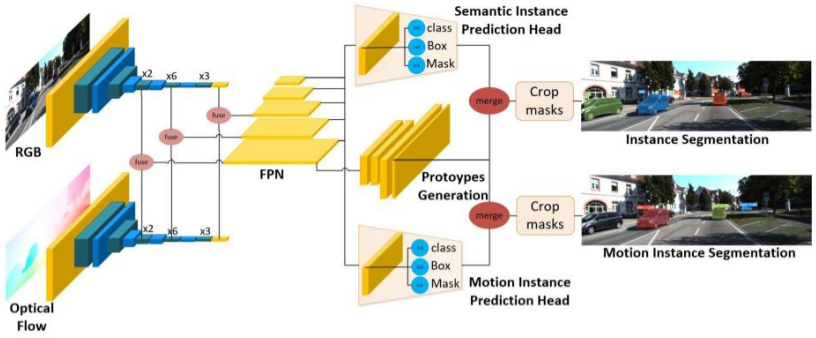

Figure 1: Overview of the proposed network architecture

Datasets	# Frames	# Sequences	# Object Categories	Instance Labels
KITTI-MoSeg [15]	1300	~ 5	1 (Car Only)	✗
KITTI-Motion [16]	455	-	1 (Car Only)	✗
KITTI-MoSeg Extended [10]	12919	~ 38	1 (Car Only)	✗
Cityscapes Motion [16]	3475	-	1 (Car Only)	✓
KITTI Instance-MoSeg (Ours)	12919	~ 38	5	✓

Table 1: Comparison of different datasets for motion segmentation.

dataset that has 12919 images with instance-wise motion masks for different classes instead of car only which is crucial for learning a generalizable class agnostic segmentation head.

2 Proposed Method

2.1 KITTI-Instance-MoSeg Dataset

Motion segmentation in autonomous driving has limited datasets [16, 15] which are primarily focused on cars. We extend KITTIMoSeg dataset provided by Rashed et al. [10] with motion and semantic instance segmentation masks for 5 classes including car, pedestrian, bicycle, truck, bus instead of the original annotations for car class only. The higher variability in the classes prevents our class agnostic head from overfitting to a certain semantic class. Table 1 shows a summary comparison of the different datasets with respect to ours.

2.2 Multi-task Learning of Semantics and Class Agnostic Instances

Our baseline model from YOLACT [1] has a feature extractor network, a feature pyramid network [5] and a protonet module that learns to generate prototypes. We experiment with different feature extractor networks as shown in section 3. In parallel to protonet, a prediction head is learned that outputs bounding boxes, classes and coefficients that are linearly combined with the prototypes to predict the instance masks. We train this model for the class agnostic (motion) instance segmentation task to serve as a baseline for comparative purposes. We refer to this baseline model *RGB-Only* since the input modality is only appearance information.

We feed the model with motion information through the Optical Flow (OF) in sintel color wheel representation. In this approach, we make use of FlowNet 2.0 [4] model to compute optical flow. It is worth noting that standard automotive embedded platforms including Nvidia Xavier, TI TDA4x and Renesas V3H have a hardware accelerator for computing dense optical flow and it can be leveraged without requiring additional processing. The fusion between appearnace (RGB) and motion (OF) is performed on the feature level we refer to this model as *RGB+OF*.

We further propose the joint model to learn semantic and class agnostic instance segmentation. Separate prediction heads D_s and D_m are used for semantic and motion instance segmentation tasks respectively as shown in Figure 1. During training, we alternate between k steps for training D_s using KITTI-Instance-MoSeg with semantic labels, and k steps for training D_m using KITTI-Instance-MoSeg motion labels or the generic DAVIS dataset. The feature extraction network, feature pyramid network and protonet are shared among the two tasks to ensure computational efficiency of our proposed multi-task learning system.

Table 2: KIIT-MoSeg Results.

Model	Moving IoU	Background IoU	mIoU	FPS
RTMotSeg [13]	50	99.1	74.6	-
FuseMODNet [10]	53.2	99.3	76.2	18
Ours	**59.7**	**99.4**	**79.5**	**39**

Table 3: DAVIS'17 Results.

Backbone	Class Agnostic		Semantic	
	Mask	Box	Mask	Box
ResNet101	23.5	**45.4**	**42.9**	**56.2**
MobileNetV2	**28.2**	45.3	**42.9**	55.6

Table 4: Comparison between different models for class agnostic (motion) instance segmentation.

Model	Backbone	FPS	#Params (M)	Time	Mask			Box		
					AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
[A] Comparison of Different Models										
RGB-Only	ResNet101	34.71	49.6	47.9	28.38	40.88	33.93	30.12	41.67	36.65
RGB+RGB	ResNet101	20.3	95.9	32.05	31.2	32.05	35.55	31.2	48.94	36
RGB+OF	ResNet101	20.3	95.9	49.2	39.26	60.02	45.76	41.24	60.59	49.28
[B] Comparison of Different Backbones										
RGB+OF	ResNet50	29.74	57.9	33.6	39.74	58.26	46.36	40.64	59.03	49.07
RGB+OF	MobileNetV2	**39.18**	12.9	**25.5**	**43.55**	68.34	**48.59**	43.93	67.73	51.56
RGB+OF	ShuffleNetV2	38	**11.1**	26.3	43.9	**68.43**	48.49	**44.9**	**68.26**	**51.77**

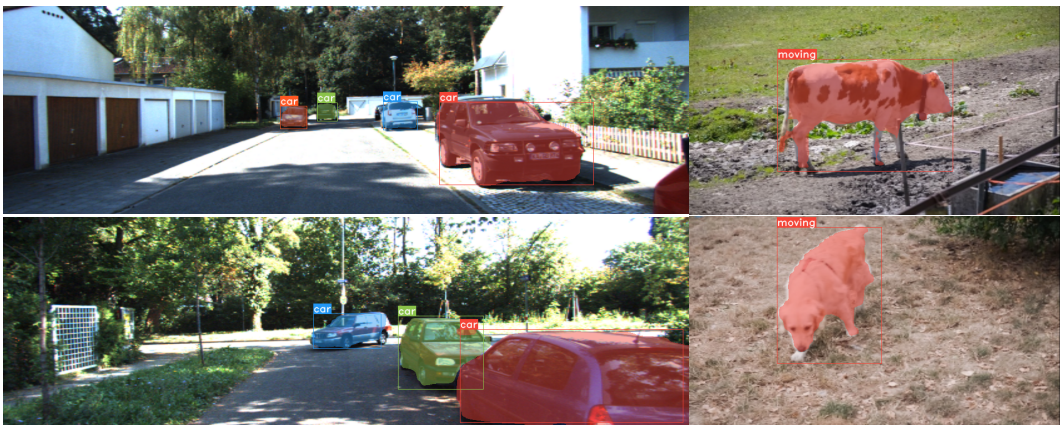

Figure 2: a) Semantic Instance Masks. b) Class Agnostic Instance Masks from Multitask Model.

3 Experimental Results

In this section, we provide the details of our experimental setup and results on KITTI-Instance-MoSeg. Our model provides 9 fps speedup over the state of the art motion segmentation methods while additionally providing instance-wise motion masks.

3.1 Experimental Setup

We train our model using SGD with momentum using an initial learning rate of 10^{-4}, momentum of 0.9 and a weight decay of $5x10^{-4}$ with batch size 4 for 150 number of epochs. A learning rate scheduling is used where it is divided by 10 at iterations 280k and 600k. Finally, we report frame rate and time in milliseconds for all models running on Titan Xp GPU on image resolution 550×550.

3.2 Benchmarking to SOA Motion Segmentation and Analysis

We benchmark our model with other SOA motion segmentation methods [14][10] in Table 2. Since we are the first to propose instance motion segmentation in autonomous driving literature we post-process the instance motion masks into pixel-level motion segmentation and compare using mean intersection over union for moving pixels and frame rate. Our model outperforms these methods in terms of mIoU while being more efficient in terms of frames per second.

Furthermore, we provide exhaustive ablation studies in Table 4. Two main factors are studied namely: (A) The impact of different input modalities to the model. (B) The impact of various backboness [3][3][6][12] to estimate the optimal performance in terms of accuracy vs speed trade-off.

Table 5: Multi-task model results on KITTI-Instance-MoSeg.

Model	Backbone	Semantic						Class Agnostic					
		Mask			Box			Mask			Box		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Multi-Task	ResNet101	22.5	39.5	22.2	27.7	54.5	25.7	44.3	66.7	50.4	44.8	66.8	55.1
Multi-Task	MobileNetV2	21.9	36.5	22.6	26.5	53.6	24.2	**45.3**	**69.0**	**57.4**	**44.3**	**69.4**	**52.4**

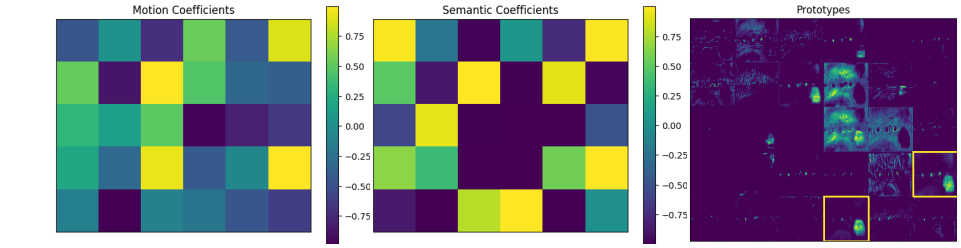

Figure 3: Prototypes and Coefficients Analysis. First row shows prototypes and coefficients for motion and semantic. Last row shows the predicted motion and instance masks.

Table 4 demonstrates quantitative results evaluated on KITTI-Instance-MoSeg dataset using various network architectures. Significant improvement of 11% in mean average precision has been observed with feature-fusion which confirms the conclusions of [15, 17, 9]. MobileNetV2 has shown to outperform other backbones while maintaining the second best mAP.

In order to assess the power of class agnostic segmentation on more general moving objects that are outside the labels within KITTIMoSeg we report results on alternate training the model between DAVIS'17 motion and KITTIMoSeg semantic annotations in Table 3. Finally, Table 5 demonstrates our results for the multi-task model with two backbones which can be compared with corresponding baseline models from Table 4. Our multi-task learning system can suffer from degradation of semantic segmentation task that can be remedied by learning a re-weighting for the losses or by training the class agnostic head while freezing the weights for the rest of the model. Nonetheless the scope of the current work is to show the efficiency of sharing protonet among these two tasks where our multi-task model runs at 34 fps while our baseline runs at 39 fps. Figure 2 further shows the multi-task model inferring semantic labels on KITTI, while still being able to segment unknown moving objects that exist in DAVIS dataset.

In order to better understand the model output, we perform an analysis on the common prototypes and coefficients learned for both motion and semantic instance segmentation. Figure 3 shows the output for the basis learned with a total of 32 prototypes in (c). The output basis are organized in a 6×5 grid which correspond to the 6×5 grid for the coefficients. Both semantic and motion coefficients are shown in (a, b), where they can be negative or positive which can help to mask or add certain objects to the final mask. The predicted final semantic and motion instance masks for the corresponding two frames are shown last row which are constructed as a linear combination using these basis and coefficients. The output shows meaningful learned basis that can help in constructing the final masks and explains what motion and semantic heads are learning.

4 Conclusions

In this paper, we developed the first class agnostic and semantic instance segmentation model for autonomous driving and provided KITTI-Instance-MoSeg dataset for these tasks. We designed a computationally efficient multi-task model for semantic and class agnostic instance segmentation through sharing protonet and learning different prototype coefficients which runs at 34 fps.

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9157–9166, 2019.

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

[3] 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 (CVPR)*, pages 770–778, 2016.

[4] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655, 2016.

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[6] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.

[7] Stefan Milz, Georg Arbeiter, Christian Witt, Bassam Abdallah, and Senthil Yogamani. Visual slam for automated driving: Exploring the applications of deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 247–257, 2018.

[8] Mohamed Ramzy, Hazem Rashed, Ahmad El Sallab, and Senthil Yogamani. Rst-modnet: Real-time spatio-temporal moving object detection for autonomous driving. *arXiv preprint arXiv:1912.00438*, 2019.

[9] Hazem Rashed, Ahmad El Sallab, Senthil Yogamani, and Mohamed ElHelw. Motion and depth augmented semantic segmentation for autonomous navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[10] Hazem Rashed, Mohamed Ramzy, Victor Vaquero, Ahmad El Sallab, Ganesh Sistu, and Senthil Yogamani. Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[11] B Ravi Kiran, Luis Roldao, Benat Irastorza, Renzo Verastegui, Sebastian Suss, Senthil Yogamani, Victor Talpaert, Alexandre Lepoutre, and Guillaume Trehard. Real-time dynamic object detection for autonomous driving using prior 3d-maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[13] Mennatullah Siam, Sara Eikerdawy, Mostafa Gamal, Moemen Abdel-Razek, Martin Jagersand, and Hong Zhang. Real-time segmentation with appearance, motion and geometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5793–5800. IEEE, 2018.

[14] Mennatullah Siam, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, and Martin Jagersand. Rtseg: Real-time semantic segmentation comparative study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1603–1607. IEEE, 2018.

[15] 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 *Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2859–2864, 2018.

[16] Johan Vertens, Abhinav Valada, and Wolfram Burgard. Smsnet: Semantic motion segmentation using deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, Canada, 2017.

[17] Marie Yahiaoui, Hazem Rashed, Letizia Mariotti, Ganesh Sistu, Ian Clancy, Lucie Yahiaoui, Varun Ravi Kumar, and Senthil Yogamani. Fisheyemodnet: Moving object detection on surround-view cameras for autonomous driving. *arXiv preprint arXiv:1908.11789*, 2019.