RST-MODNet: Real-time Spatio-temporal Moving Object Detection for Autonomous Driving

Mohamed Ramzy[1], **Hazem Rashed**[2], **Ahmad El Sallab**[2] **and Senthil Yogamani**[3]
[1]Cairo University [2]Valeo R&D, Egypt [3]Valeo Vision Systems, Ireland
mohamed.ibrahim98@eng-st.cu.edu.eg
{hazem.rashed,ahmad.el-sallab,senthil.yogamani}@valeo.com

Abstract

Moving Object Detection (MOD) is a critical task for autonomous vehicles as moving objects represent higher collision risk than static ones. The trajectory of the ego-vehicle is planned based on the future states of detected moving objects. It is quite challenging as the ego-motion has to be modelled and compensated to be able to understand the motion of the surrounding objects. In this work, we propose a real-time end-to-end CNN architecture for MOD utilizing spatio-temporal context to improve robustness. We construct a novel time-aware architecture exploiting temporal motion information embedded within sequential images in addition to explicit motion maps using optical flow images. We demonstrate the impact of our algorithm on KITTI dataset where we obtain an improvement of 8% relative to the baselines. We compare our algorithm with state-of-the-art methods and achieve competitive results on KITTI-Motion dataset in terms of accuracy at three times better run-time. The proposed algorithm runs at 23 fps on a standard desktop GPU targeting deployment on embedded platforms.

1 Introduction

The Autonomous Driving (AD) scenes are highly dynamic as they contain multiple object classes that move at different speeds in diverse directions [2, 3]. It is critical to understand the motion model of each of the surrounding elements for the purpose of planning the ego-trajectories considering the future positions and velocities of these objects. Typically, there are two types of motion in a an autonomous driving scene, namely motion of surrounding obstacles and the motion of the ego-vehicle. It is challenging to successfully classify the surrounding objects as moving or static when the camera reference itself is moving. In this case, even the objects that are not moving will be perceived as dynamic ones. Moving object detection implies two tasks that are performed jointly, namely, generic object detection which extracts specific classes such as pedestrians and vehicles. This is followed by motion classification, in which a classifier identifies the motion state of the object at hand among two classes, dynamic and static. Object detection and semantic segmentation has become a mature algorithm for automated driving [4] but motion segmentation is relatively an unexplored problem. Recent automated driving datasets [5] include moving object detection task.

Recent CNN-based algorithms [6, 7, 8] explored the problem of end-to-end motion segmentation through usage of optical flow images providing the motion of the surrounding scene as a prior information for the network which learns to generate a binary mask of two classes, "Moving" and "Static". Motion segmentation can be integrated into a multi-task system along with segmentation and other tasks [9, 10]. Motion Estimation also helps in better estimation of depth [11]. Motion information can be perceived implicitly through a stack of temporally sequential images [12], or explicitly through an external motion map such as optical flow map[13]. Implicit motion modelling

Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

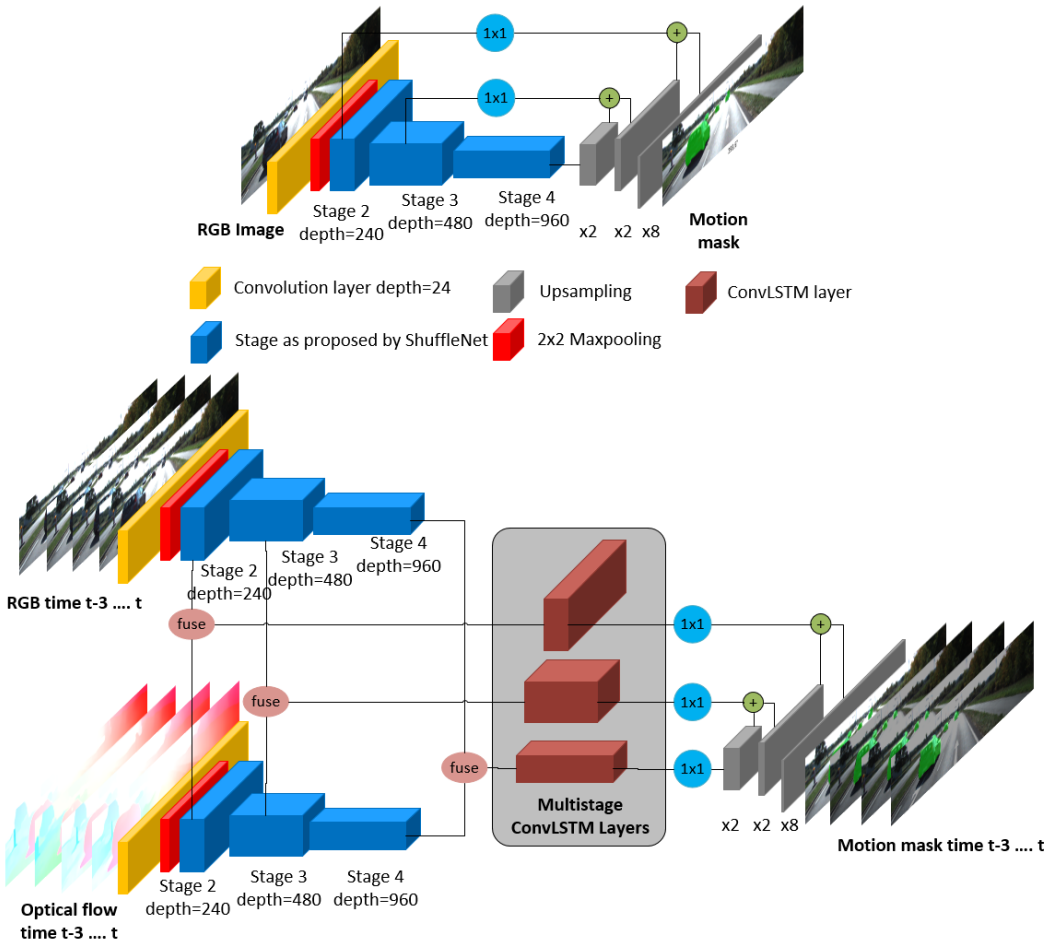

Figure 1: **Top:** Our baseline architecture based on ShuffleNet [1]. **Bottom:** Our proposed architecture after optical flow and time augmentation.

is prone to failure due to increased complexity of the task as the network learns to model motion in addition to segmenting the interesting object classes. On the other hand, external optical flow encodes motion between two consecutive frames only without considering previous states of the scene which negatively affects the output in two ways. First, the network becomes sensitive to optical flow errors because motion is being understood from two frames only. Second, the output masks become temporally inconsistent as they are independent of each other across time and therefore masks are more prone to errors and rapid changing. Moreover, optical flow encodes two pieces of information, the motion of the surrounding obstacles, in addition to the motion of the ego-vehicle which results in significant motion vectors associated with the static objects as well. This leads to the incorrect perception of static objects as moving objects. Nevertheless, optical flow augmentation has proven to improve accuracy of MOD compared to motion modelling from single color images due to understanding the motion across two sequential images such as in [14, 6, 15]. These results raised our question of how a CNN architecture would behave if it considers the previous states of the surrounding obstacles.

In this work, we explore the benefit of leveraging temporal information through implementation of time-aware CNN architecture in addition to explicit motion modelling through optical flow maps. Moreover, we focus on real-time performance due to the nature of the autonomous driving task [16, 17, 7]. To summarize, the contributions of this work include:

- Implementation of a novel CNN architecture for MOD utilizing spatio-temporal information. Our model combines both explicit and implicit motion modelling for maximum performance, and unlike previous baselines it ensures temporal consistency between successive frames.

- Construction of real-time performance network which significantly outperforms state-of-the art approaches and becomes suitable for time-critical applications such as the autonomous driving.
- Ablation study for various time-aware architectures for maximum performance in terms of accuracy, speed and temporal consistency.

The rest of the paper is organized as follows: a review of the related work is presented in Section 2. Our methodology including the dataset preparation and the used network architectures is detailed in Section 3. Experimental setup and final results are illustrated in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

Classical approaches based on geometrical understanding of the scene such as [18] have been suggested for motion masks estimation. Wehrwein et al. [19] introduced some assumptions to model the motion of the background as homography. This approach is very difficult to be used in AD due to the limited assumptions which causes errors such as camera translations assumptions. Classical methods generally provide less performance than deep learning methods in addition to the need to use complicated pipelines which introduce higher complexity in the algorithm. For instance, Menze et al. [18] runs at 50 minutes per frame which is not suitable for AD.

Generic foreground segmentation using optical flow has been proposed by Jain et. al.[20], however it does not provide information about the state of each obstacle whether it is moving or static. In [21, 22] video object segmentation has been studied, however these networks are not practical for AD due to high complexity where they depend on R-CNN as in [21], and DeepLab as in [22] which run in 8 fps. Motion segmentation using deep network architectures has been explored by Siam et al. [6, 15]. These networks rely only on explicit motion information from optical flow which makes them sensitive to the optical flow estimation errors. Fisheye MOD has been studied in [12] using publicly available fisheye dataset [5] proving the importance temporally sequential images in MOD.

LiDAR sensors have been explored for MOD as as well, where most of LiDAR-based methods used clustering approaches such as [23] to predict the motion of points using methods such as RANSAC, and then clustering takes place on the object level. Deep learning has been explored as well for such problem. 3D convolution is used in [24] to detect vehicles. Other methods projected the 3D points on images to make use of 2D convolutions on the image 2D space instead of 3D space [25]. Low-illumination MOD has been explored by [26] where optical flow has been utilized from both camera and LiDAR sensors demonstrating the importance of explicit motion modelling. Recent work [27] predicts motion of objects from two input Lidar scans. This method uses implicit learning for motion information through two sequential Lidar scans and does not discuss the impact of time-aware networks which motivates our work to towards this study.

In this work, we aim to provide a real-time cost-effective approach for the autonomous driving application. Unlike LiDAR sensors, camera sensors have high efficiency compared to their cost. Thus, we focus on camera-based motion segmentation approaches.

3 Methodology

In this section we discuss dataset preparation, and detail the proposed architectures for our experiments.

3.1 Dataset Preparation

3.1.1 Annotations Generation

To be able to train our deep model for good generalization, a large dataset including motion masks annotations is needed. There is huge limitation in MOD public datasets. For instance, Siam et al. [6] provides 1300 images only with weak annotation for MOD task. On the other hand, 255 annotated frames are provided by Valada et al. [14] on KITTI dataset, while 3475 frames are provided on Cityscapes [28]. Cityscapes does not provide information about 3D motion which makes it not

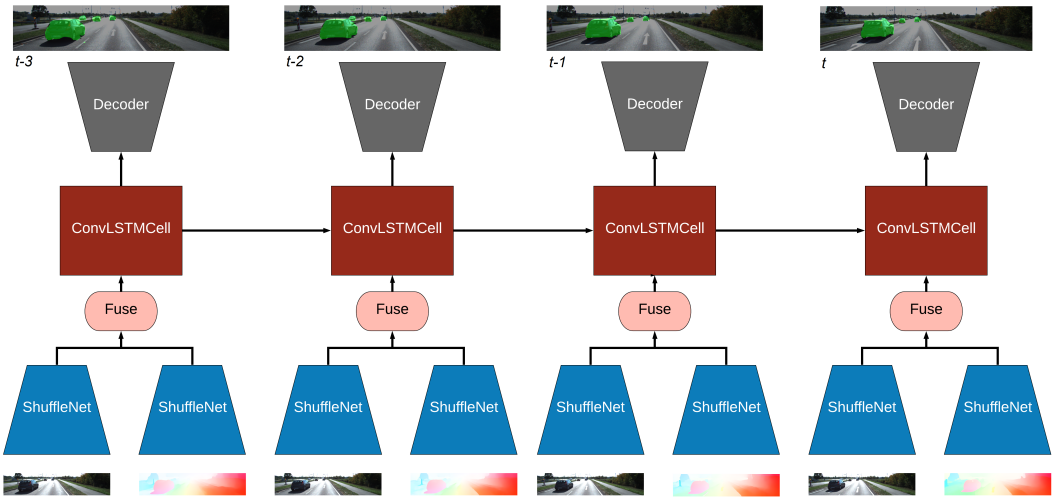

Figure 2: Detailed architecture for our approach demonstrating how temporal information is used within the unfolded ConvLSTM cell.

suitable for us to extend the dataset. Behley et al. [29] provides MOD annotations for 3D point clouds only, but not for dense pixels. Due to this limitation in datasets, we build our own MOD dataset. We make use of the method in [6] to generate motion masks from KITTI in order to extend the KittiMoSeg dataset. The bounding boxes provided by KITTI dataset are projected on 2D images while the tracking information is used 3D velodyne coordinate system to compute the velocity of each object compared to the ego-vehicle. The ego-vehicle velocity is provided via GPS annotation which allows us to compute relative speed between the ego-vehicle and the surrounding objects. We compare the relative speed to predefined thresholds to classify whether the obstacle is moving or static. This method is used to generate semi-automatic annotation for 12k images and then manual filtering is performed for fine tuning.

3.1.2 Color Signal

KITTI dataset[30] is used as it provides temporal sequences which we exploit to develop our time-aware architecture. Moreover, it has 3D bounding box and GPS coordinates annotations which we use to build our motion mask annotations. The upper part of the image is cropped as it is mainly dominated by sky and has no information about moving objects. The final resolution we train our network with is 256x1224.

3.1.3 Motion Signal

Motion can be either implicitly learned from temporally sequential frames, or provided explicitly to the system through an input motion map, as for example optical flow maps. In our approach, we develop a hybrid model that combines both methods together for maximum performance. FlowNet[31] architecture is leveraged to generate the flow maps. We use color-wheel representation as it was observed to provide the best performance for encoding both magnitude and direction of motion. This is consistent with the observation of [6, 32].

3.2 Network Architecture

In this section, we detail our baseline architecture, discuss how to use motion information and how to maximize the benefit of temporal sequences for MOD.

3.2.1 Baseline Architecture

Our baseline model is based on [33]. The network is composed of an encoder-decoder architecture where feature extraction is performed by [1] reducing computational cost at high level of accuracy which is perfect for AD application. The decoder is based on [34] which consists of 3 transposed

Table 1: Quantitative comparison between different network architectures for MOD.

Experiment	mIoU	Moving IoU
RGB	65.6	32.7
RGB+Flow	74.24	49.36
RGB+Flow frame stacking	63	27
RGB+Flow 3D Convolution	64.3	29.8
RGB+Flow - LSTM (Early)	73.5	48
RGB+Flow - LSTM (Late)	69.2	39.3
RGB+Flow - LSTM (Multistage-2-filters)	73.7	48.5
RGB+Flow - GRU (Multistage)	75	50.9
RGB+Flow - LSTM (Multistage)	**76.3**	**53.3**

convolution layers that are used to upsample the low resolution feature maps to the original image resolution. The network is trained to predict two classes, i.e, Moving and Non-Moving. There is huge imbalance between the two classes because of the background pixels which are considered static pixels as well. Weighted cross entropy is used to tackle the problem. The baseline architecture is used to evaluate the impact of RGB images only on MOD as illustrated in Figure 1

3.2.2 Motion Augmentation

As demonstrated by [6], explicit motion modelling through optical flow provides more accuracy than implicit learning of motion cues through temporal sequences. This is done through a 2-stream mid-fusion architecture which combines the processed feature maps from both network branches. It has been shown by [13, 32] that mid-fusion architecture outperforms early-fusion which is based on raw data fusion before feature extraction. Feature-level fusion provides maximum accuracy at the cost of network complexity as the number of weights in encoder part is doubled. We adopt this approach for comparative study where semantic information is combined with motion information as illustrated in Figure 1 and we demonstrate the impact on real-time performance.

3.2.3 Time Augmentation

The main contribution of this work is to study the impact of including temporal information for MOD. For that purpose, we build upon the mid-fusion network and provide empirical study for various time-aware network architectures. We discuss the effect of using Frame stacking, 3D convolution, ConvLSTM[35], and simpler GRU[36] layers which are time-dependent. For such experiments, we use a batch of 4 images as input as illustrated in Figure 2 which explains how ConvLSTM is unfolded utilizing sequence of images for MOD. We design three network architectures leveraging ConvLSTM layers and provide empirical study to demonstrate their benefit for MOD.

Early-LSTM: In this case, we refer to the usage of ConvLSTM layer in each encoder separately, then fusion is done on the processed information.

Late-LSTM: In this case, we refer to the usage of ConvLSTM at the decision level before softmax layer where the network learns to use time information before the final classification is done.

Multistage-LSTM: We implement several ConvLSTM layers across the network at 3 different stages as illustrated in Figure 1 (Bottom). Finally, by "Multistage-2-filters" we refer to using 1x1 convolutional layers which squeezes the depth of the feature maps to *num_classes* and then apply ConvLSTM to the output channels.

4 Experiments

4.1 Experimental Setup

In our experiments, ShuffleSeg [33] model was used with pre-trained ShuffleNet encoder on Cityscapes dataset for semantic segmentation except for the 3D Convolution experiment as we randomly initialized the encoder weights. For the decoder part, FCN8s decoder has been utilized with

Table 2: Quantitative results on KITTI-Motion[14] dataset in terms of mean intersection over union (mIoU) and running frames per second (fps) compared to state-of-the-art methods.

Experiment	mIoU	fps
CRF-M[37]	77.9	0.004
MODNet[6]	72	6
SmSNet[14]	**84.1**	7
RTMotSeg[15]	68.8	**25**
RST-MODNet-GRU (ours)	82.5	23
RST-MODNet-LSTM (ours)	83.7	21

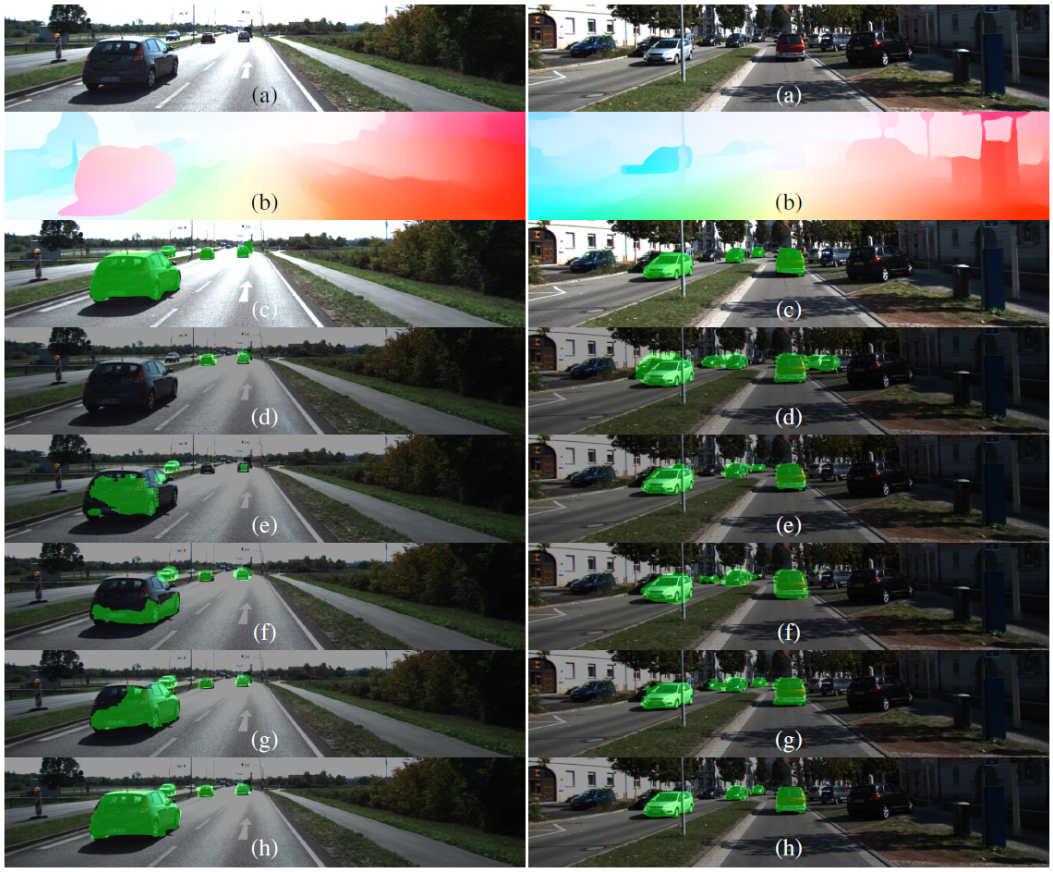

Figure 3: Qualitative comparison of our algorithm on two samples of KITTI dataset each sample is represented in a column. **(a),(b)** show the input RGB images and flow images. **(c)** shows ground truth. **(d)** shows RGB-only output. **(e)** shows RGB+Flow output. **(f)** output of RGB+Flow-LSTM(Late). **(g)** shows output of RGB+Flow-GRU(Multistage). **(h)** shows output of the proposed architecture RGB+Flow-LSTM(Multistage).

randomly initialized weights. L2 regularization with weight decay rate of $5e^{-4}$ and Batch Normalization are incorporated. We trained all our models end-to-end with weighted binary cross-entropy loss for 200 epochs using 4 sequential frames. Adam optimizer is used with learning rate of $1e^{-4}$. For frame stacking experiments we modified the depth of the first convolutional layer filters to match the input by replicating the filters in the depth dimension to utilize Cityscapes weights instead of randomly initializing them.

4.2 Experimental Results

We provide a table of quantitative results for our approach evaluated on KITTI dataset and a table for comparison with state-of-the-art baselines on KITTI-Motion dataset [14]. Qualitative evaluation is illustrated in Figure 3. Table 1 demonstrates our results in terms of mIoU on both classes in addition to IoU for the Moving class. RGB-only experiment result is used as a comparison reference where color information only is used for classifying MOD without using either implicit or explicit motion modelling. Significant improvement for 17% in moving class IoU has been observed after fusion with optical flow, which is consistent with previous conclusions in [6, 15]. Naive frame stacking showed inability of the network to benefit from the temporal information embedded into the sequence of images while 3D convolution increased the network complexity dramatically which made it not suitable for embedded platform for autonomous driving application. For that reason we focus our experiments on usage of ConvLSTM layers where we provide an empirical evaluation to figure out the best architecture utilizing ConvLSTM. Three architectures are implemented using ConvLSTM. Early and Late LSTM show improved performance over the baseline RGB, however they perform very close to standard two-stream RGB+Flow which means the information is not fully utilized. This encourages the implementation of a mid-level ConvLSTM-based network that captures motion in multiple stages. Using this architecture, we obtain absolute improvement of 4% in accuracy and relative improvement of 8% over RGB+Flow showing the benefit of our approach. We provide two versions of our multistage architecture comparing ConvLSTM and GRU. We observe very close performance in terms of accuracy with slightly higher running rate using GRU which is expected due to simpler internal architecture.

Table 2 shows a comparison between our approach and state-of-the-art baseline methods. For fair comparison of model speed, we run all the tests on our Titan X Pascal GPU using the same input resolution in [14]. RTMotSeg[15] has two models, one of which is using LiDAR point cloud in a post processing step to minimize false positives. We report the model which does not use LiDAR sensor as we mainly focus on camera-based solutions. It is shown that our method is on par with the baseline methods where we provide almost the same accuracy as SMSNet[14], however at almost double the inference speed using both our multistage time-aware architectures which makes them more suitable for embedded platform for autonomous driving applications.

Figure 3 shows qualitative results on two KITTI samples demonstrating the benefit of using time-aware networks for MOD where each column represents a sample. (a),(b) show the input RGB and optical flow inputs. (c) shows the motion mask ground truth. (d) shows inability of CNN to understand motion information from color images only without sequence of images or optical flow maps. (e) shows improvement over RGB-only due to fusion with optical flow which encodes motion of the scene. (f) shows the output of RGB+Flow after adding LSTM layer before softmax layer (Late) which demonstrates the improvement over RGB-only as illustrated in Table 1. However, the network is still unable to completely utilize the motion information embedded within the sequence of images. (g),(h) show the output of our multistage models,namely RGB+Flow-GRU in (g) and RGB+Flow-LSTM in (h). Results visually confirm the benefit of our algorithm through implementation of multistage time-aware layers where motion information is fully exploited.

Figures 4,5 show the advantage of our approach across time where relationship between sequential images has been modelled within the network. Each figure represents a sample sequence from KITTI data where the left column represents the output of RGB+Flow while the right column shows the impact of our algorithm considering time information. On the left column, the furthest car in time *t-3* has been segmented correctly then accuracy is lost in *t-2* and obtained again in *t-1*. This is also shown on the close car on the left where the mask is sensitive to optical flow map. On the other hand, the right column shows temporally consistent motion masks after the addition of our multistage-LSTM layers within the network. The same conclusion is obtained from Figure 5, where these results demonstrate the improved performance in Table 1.

5 Conclusions

In this paper, we propose a novel method for moving object detection which balances between high accuracy and high computational efficiency. Our proposed method exploits both external motion modelling and time-aware architectures to maximize benefit from temporal motion information. An ablation study is provided for various time-aware architectures to evaluate the impact of our approach

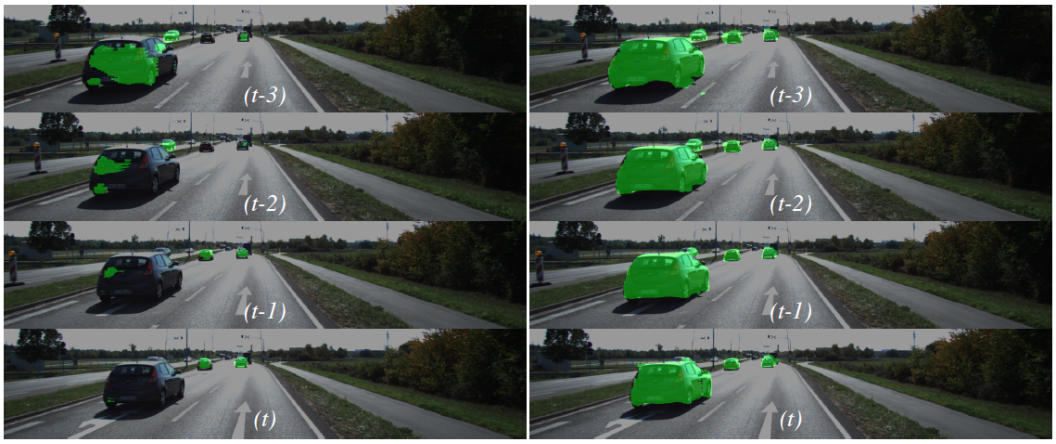

Figure 4: Qualitative evaluation demonstrating the temporal consistency obtained from our approach on the right column compared to RGB+Flow on the left column as previous baselines.

Figure 5: Qualitative evaluation demonstrating the temporal consistency obtained from our approach on the right column compared to RGB+Flow on the left column as previous baselines.

on MOD. The algorithm is evaluated on KITTI and KITTI-Motion datasets against state-of-the-art baselines. We obtain 8% relative improvement in accuracy after augmentation of time-aware layers. Competitive results are demonstrated in terms of accuracy compared to state-of-the-art SMSNet model at three times the inference speed which makes our algorithm more suitable for autonomous driving application.

References

[1] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[2] Jonathan Horgan, Ciarán Hughes, John McDonald, and Senthil Yogamani. Vision-based driver assistance systems: Survey, taxonomy and advances. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 2032–2039. IEEE, 2015.

[3] Markus Heimberger, Jonathan Horgan, Ciarán Hughes, John McDonald, and Senthil Yogamani. Computer vision in automated parking systems: Design, implementation and challenges. *Image and Vision Computing*, 68:88–101, 2017.

[4] Mennatullah Siam, Sara Elkerdawy, Martin Jagersand, and Senthil Yogamani. Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017.

[5] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. *arXiv preprint arXiv:1905.01489*, 2019.

[6] Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand, and Ahmad El-Sallab. Modnet: Motion and appearance based moving object detection network for autonomous driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2859–2864. IEEE, 2018.

[7] 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.

[8] 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.

[9] Ganesh Sistu, Isabelle Leang, Sumanth Chennupati, Stefan Milz, Senthil Yogamani, and Samir Rawashdeh. Neurall: Towards a unified model for visual perception in automated driving. *arXiv preprint arXiv:1902.03589*, 2019.

[10] Sumanth Chennupati, Ganesh Sistu, Senthil Yogamani, and Samir A Rawashdeh. Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[11] Varun Ravi Kumar, Stefan Milz, Christian Witt, Martin Simon, Karl Amende, Johannes Petzold, Senthil Yogamani, and Timo Pech. Monocular fisheye camera depth estimation using sparse lidar supervision. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2853–2858. IEEE, 2018.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] B Ravi Kiran, KP Anoop, and Y Senthil Kumar. Parallelizing connectivity-based image processing operators in a multi-core environment. In *2011 International Conference on Communications and Signal Processing*, pages 221–223. IEEE, 2011.

[17] Alexandre Briot, Prashanth Viswanath, and Senthil Yogamani. Analysis of efficient cnn design techniques for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 663–672, 2018.

[18] 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.

[19] Scott Wehrwein and Richard Szeliski. Video segmentation with background motion models. In *BMVC*, volume 245, page 246, 2017.

[20] 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 *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2126. IEEE, 2017.

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

[22] 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.

[23] Ayush Dewan, Tim Caselitz, Gian Diego Tipaldi, and Wolfram Burgard. Motion-based detection and tracking in 3d lidar scans. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4508–4513. IEEE, 2016.

[24] Bo Li. 3d fully convolutional network for vehicle detection in point cloud. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1513–1518. IEEE, 2017.

[25] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3d lidar using fully convolutional network. *arXiv preprint arXiv:1608.07916*, 2016.

[26] 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.

[27] Ayush Dewan, Gabriel L Oliveira, and Wolfram Burgard. Deep semantic classification for 3d lidar data. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3544–3549. IEEE, 2017.

[28] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[29] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

[30] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[31] 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.

[32] Hazem Rashed, Senthil Yogamani, Ahmad El Sallab, Pavel Krízek, and Mohamed El-Helw. Optical flow augmented semantic segmentation networks for automated driving. In *VISIGRAPP*, 2019.

[33] Mostafa Gamal, Mennatullah Siam, and Moemen Abdel-Razek. ShuffleSeg: Real-time semantic segmentation network. *arXiv preprint arXiv:1803.03816*, 2018.

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[35] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[36] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[37] N Dinesh Reddy, Prateek Singhal, and K Madhava Krishna. Semantic motion segmentation using dense crf formulation. *arXiv preprint arXiv:1504.06587*, 2015.