Enhancement of Visual SLAM Precision Using Semantic Image Segmentation for Automotive Systems

Patrick Rebling[1], Michael Weber[1,2], Tobias Weiß[1,2], Franck Gechter[2], Reiner Kriesten[1], and Philipp Nenninger[1]

[1] Institute of Energy Efficient Mobility (IEEM), Karlsruhe University of Applied Sciences
{patrick.rebling, michael.weber, tobias.weiss, reiner.kriesten, philipp.nenninger}@h-ka.de
[2] CIAD (UMR 7533), Université Bourgogne Franche-Comte, UTBM
LORIA-MOSEL (UMR 7503), Université de Lorraine
franck.gechter@utbm.fr

Abstract. In the field of advanced driver assistance systems (ADAS) testing and autonomous driving (AD) feature evaluation, novel approaches relying on augmented reality (AR) promise to deliver cost-saving benefits. These forward-looking approaches leverage vSLAM techniques to create mapping solutions that are essential for augmentation. A critical challenge, however, is maintaining the high precision required for these maps and, by extension, the SLAM algorithm itself. This precision is often compromised by the presence of false-positive detections of feature points. In response to this challenge, this paper presents an improvement to the ORB-SLAM3 algorithm. The proposed approach incorporates semantic segmentation without compromising processing speed to increase the precision and reliability of the SLAM system. This is to ensure that the integration of AR-based solutions in the automotive sector is both effective and sustainable, providing tangible benefits in the testing and development of ADAS and autonomous driving technologies.

Keywords: Artificial Intelligence, Augmented Reality, Advanced Driver Assistance Systems, Visual Simultaneous Localization and Mapping, 3-Dimensional Modeling, Image Segmentation, Object Detection

1 Introduction and Motivation

Understanding of the surrounding world of vehicles and robots is essential for automation [1, 2] and testing of automated systems using augmented reality (AR) [3]. While there are approaches that only use the current perception of the vehicle for planning , others are based on pre-built maps for localization and orientation. These maps are required for automated driving to interpret and analyse the current state of the environment and to navigate the vehicle efficiently. A common approach for map generation is the usage of *Light Detection and Ranging* (Lidar) which allows robust creating of maps and the relocalization inside those maps but the sensors are quite expensive. Pure visual approaches require only cameras for mapping and localization and are therefore cost-efficient and suitable for automotive and robotic applications. For visual approaches, it is necessary to have knowledge about meaningful feature points for relocalization. Therefore, we present an approach for enhancing visual Simultaneous Localisation and Mapping (vSLAM) in automotive applications by filtering of feature points for gaining robustness for AR-based testing of advanced driver assistance systems (ADAS).

2 State-of-the-Art

In this section, we will focus on the state-of-the-art in mapping and localization of vehicles and information extraction from images based on semantic segmentation. After a brief overview on techniques and approaches, the algorithms chosen for our experiments will be presented and their functionality will be explained.

2.1 Simultaneous Localization and Mapping (SLAM)

Various methods have been established to map the environment only based on images in a three-dimensional virtual map like Structure From Motion (SfM), vSLAM, and Visual Odometry (VO) [4]. This offers the possibility that no sensors other than the camera system are needed to create a 3D model of the environment. Thus, there is no need for costly sensors like Lidar or Radar. While VO attempts to reconstruct the path step by step, pose by pose and may optimize only the last n poses of the trajectory using a window-based bundle fitting approach, vSLAM is primarily concerned with ensuring the consistency of the global map. The framework of vSLAM algorithms is mainly composed of three basic modules: Initialization, Tracking, Mapping, and two additional modules: Re-localization and Global Map Optimization (including Loop Closing) [5]. For the use of SLAM in automotive vehicles and the associated properties such as fast scene changes and low texturing of the environment, various approaches are available using vSLAM [6, 7]. In this paper, we focus on the *ORB-SLAM3* [8] algorithm and show how its precision can be enhanced by using semantic image segmentation.

The ORB-SLAM Algorithm. The first version of ORB-SLAM was presented in 2015 and the current version, ORB-SLAM3, supports of different camera systems and performs better than comparable SLAM algorithms [7]. Here, ORB-SLAM3 represents a complete SLAM system for monocular, stereo, and RGB-D cameras which operates in real-time and achieves remarkable results in terms of accuracy and robustness in a variety of different environments. ORB-SLAM3 is used for indoor sequences, drones, and urban driving. It consists of three main parallel threads: Tracking, Local Mapping, and Loop Closing. This algorithm is a feature-based approach, which represents the detected points in a three-dimensional *MapPoint*. Initially, the approach was to use a mono-based camera and an *Inertial Measurement Unit* (IMU) for mapping. It was noticeable that the robustness of the relocalization and the estimation of the scaling are not sufficient due to the lack of depth information. Switching to stereo-based cameras allow more accurate mapping of the environment with more realistic scaling.

2.2 Image Segmentation

Image segmentation, a fundamental process in digital image analysis, involves dividing a digital image into different segments to simplify its complexity and enable subsequent analysis of the individual segments. In this technical process, labels are assigned to individual pixels, identifying objects, people, or other relevant elements in the image [9]. An important application of image segmentation is in object recognition [10]. Instead of processing the entire image as a whole, an image segmentation algorithm is often used to identify objects of interest in the image. Then, an object detector can operate within predefined bounding boxes set by the segmentation algorithm. With this approach, the detector does not have to analyze the entire image, which increases accuracy and reduces the time required for inference. Image segmentation plays a central role in the

field of computer vision technologies and algorithms and is widely used in various practical applications. These applications include medical image analysis, computer vision for autonomous vehicles, face recognition and detection, video surveillance, and satellite image analysis [11]. While there are several methods for image segmentation, Figure 1 illustrates the three main types.

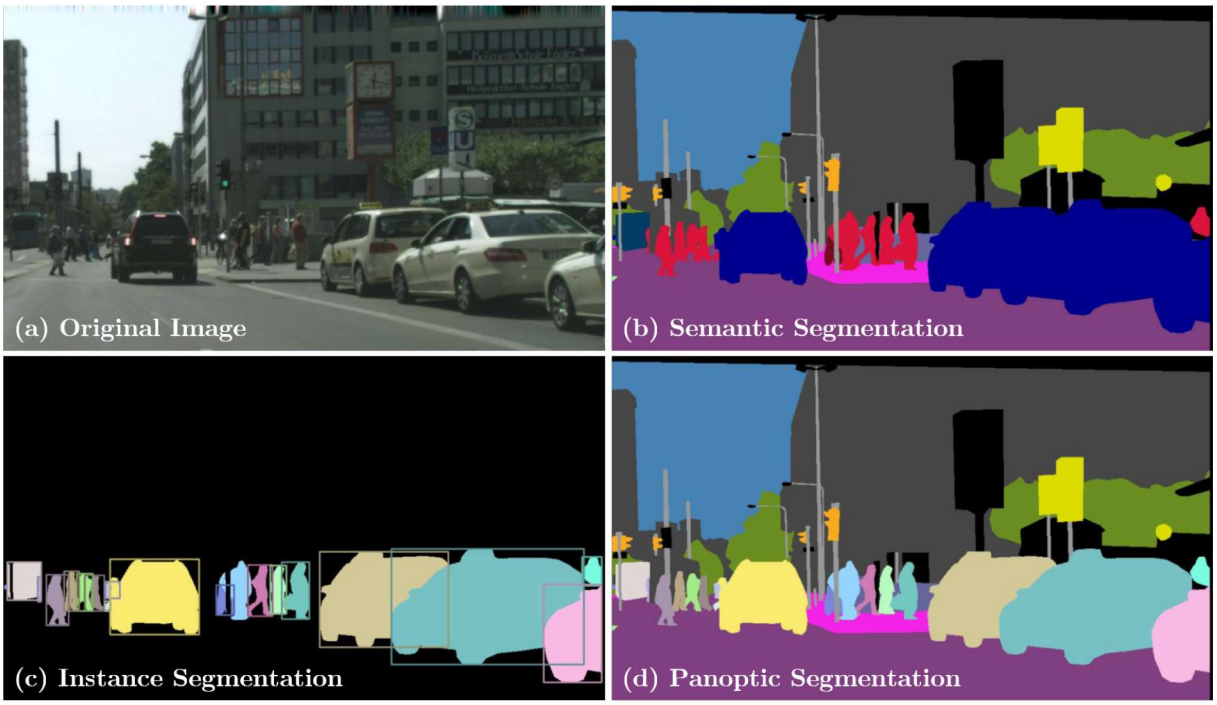


Fig. 1. Some of the main types for image segmentation. While semantic segmentation (b) does not highlight different instances of the same object type, instance segmentation (c) and panoptic segmentation (d) provides detailed information, e.g. for tracking of objects. [12, p. 1]

3 Problem Formulation

When testing ADAS through AR applications, an algorithm that can handle both mapping and tracking is a critical requirement. The challenges posed in this context are significantly more demanding than for conventional AR applications. This increased complexity arises primarily from two key factors. First, the mapping and tracking processes must exhibit exceptionally high precision, even at higher speeds. Second, these tasks must be performed within the test scenarios of the European New Car Assessment Programme (Euro NCAP) [13]. The unique test environments defined by Euro NCAP present a number of challenges, as they are very large and there are hardly any existing textures that can prevent the algorithm from finding its way. In response to this multi-faceted problem, our paper attempts to answer the following question:

> What is a Robust Simultaneous Localization And Mapping Algorithm for Mapping and Tracking to Test Camera-based Advanced Driver Assistance Systems in Low Texture European New Car Assessment Programme-Test Scenarios?

As shown on the left side in Figure 2, feature points are detected in the sky as well as on the hood of the ego vehicle, which hinder the performance of the vSLAM algorithm and causes wrong scaling of the environment inside generated maps. Experiments show that ORB-SLAM3 is struggling with the moving vehicle hood and therefore misunderstands the scales of the environment and static objects. To filter these false-positive feature points, two possible approaches are given: (1) optical flow, (2) a suitable image segmentation. As the difference between sky and distant background which provides indeed good orientation points will result in similar optical flow, we will focus on the more promising approach of image segmentation.

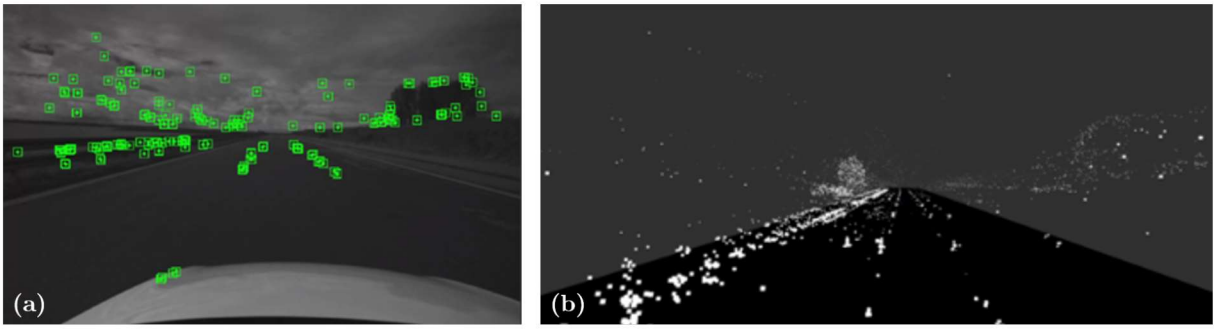


Fig. 2. ORB-SLAM3 detection results (a) and the resulting map (b) with plenty of false-positive feature points regarding the sky and the vehicle hood.

4 Approach

Our proposed solution can be divided into two core areas: 3D Global Point Cloud Map Generation, implemented using the ORB-SLAM3 framework, and the Implementation of Image Segmentation as shown in Figure 3 based on the idea of the DS-SLAM [14].

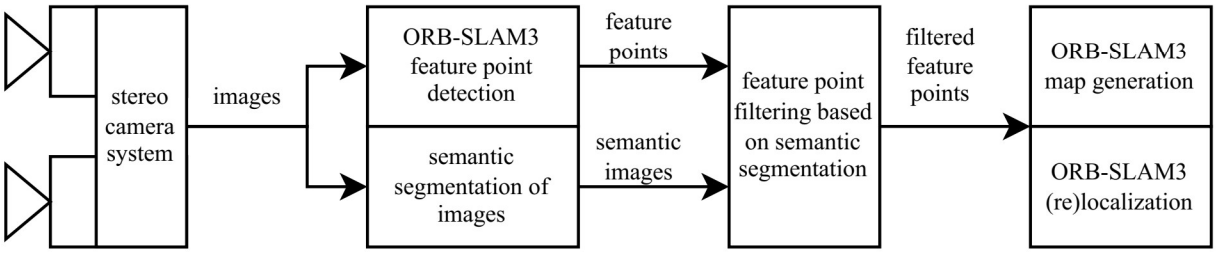


Fig. 3. Enhanced mapping pipeline of ORB-SLAM3. Parallel calculation of feature points and semantic segmentation is done on the raw images from the stereo camera. Afterwards feature points are filtered and feed into the ORB-SLAM3 map generation.

4.1 Suitable Image Segmentation Algorithm

Our initial foray into the literature revealed that we chose the *DeepLabV3+* model [15, 16]. DeepLabV3+, which uses the *Xception-65* model as its backbone [17], has great versatility in training on different datasets. The use of DeepLabV3+ trained on the *ADE20k*

dataset [18, 19] yields commendable results, accurately detecting the sky and road, albeit with some limitations, notably occasional inaccuracies in hood detection due to mirror effects and suboptimal road detection. Nevertheless, the processing speed optimized by *TensorRT* achieves a frame rate of about 30 fps, which meets the requirements of our particular use case.

Given the compelling need to correctly classify feature points within these regions, we explored alternative approaches. Xie et al. performed a comparative analysis of different neural networks for different datasets [20] shown in Figure 5. Their evaluation showed that *SegFormer* outperforms other options by providing a better balance between processing speed and accuracy, especially on the ADE20K dataset. SegFormer is shown to be robust to noise, fog, blur, snow, frost and splatter, as NVIDIA [21] points out.

With these considerations in mind, we seamlessly integrated the SegFormer B3 algorithm, as shown in Figure 4, into our semantic segmentation framework.

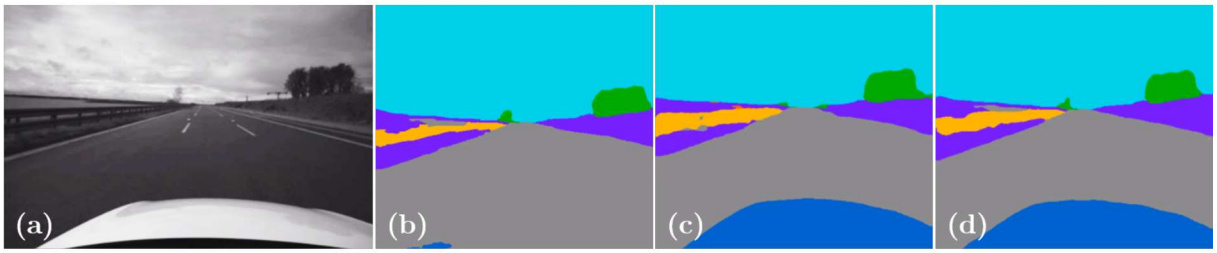


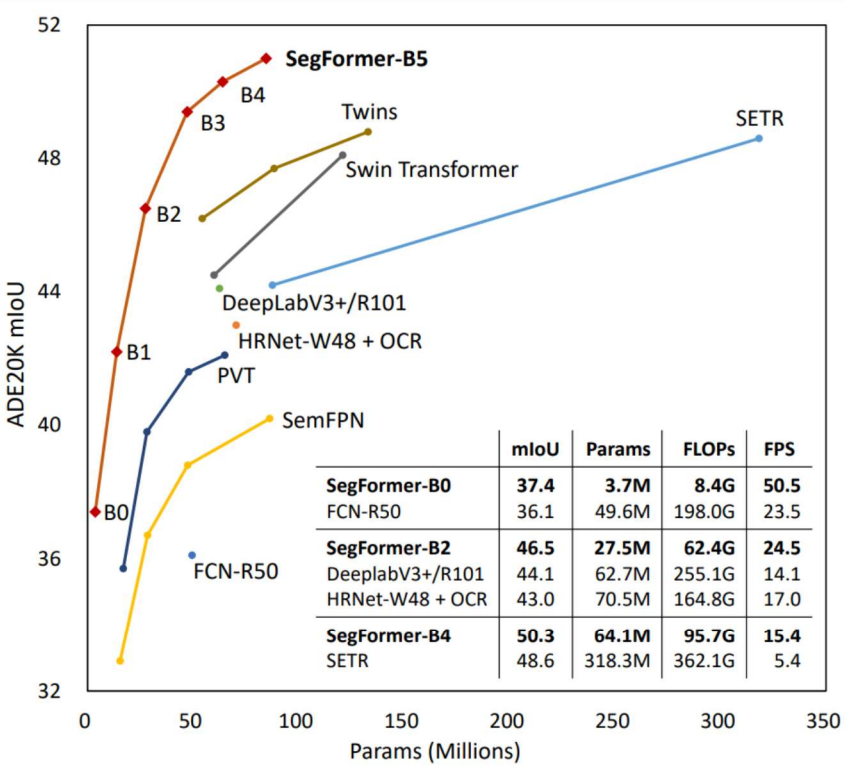
Fig. 4. Comparison between DeepLabV3+ (b), SegFormer B2 (c) and SegFormer B3 (d) on the original image (a). As shown in the figure, DeepLabV3+ is not able to detect the vehicle hood in this frame which would lead to false-positive feature points.

4.2 Feature Point Filtering

After assigning the individual pixels to their respective regions, the next phase involves processing feature points based on these region assignments. First, the region assigned to the pixel is linked to the corresponding feature point, establishing a link between the position of the feature point in the image and a particular pixel. Based on the previous assignment of the region to the pixel, this region is linked to the feature point via a corresponding index. A decision is then made as to whether the feature point is in a region that should be excluded from further use, or whether it falls in a region where feature points should be retained. At this point, additional information is assigned to each feature point, which is specified with the values -1, 0 or 1, depending on the region and the corresponding index.

$$FeaturePoint = \begin{cases} -1, \text{ for Region } \textit{Sky} \\ -1, \text{ for Region } \textit{Hood} \\ 1, \quad \text{for Region } \textit{Street} \\ 0, \quad \text{for other Regions} \end{cases}$$

In particular, we assign specific indices to the feature points detected on the road, a crucial aspect for the subsequent phase of our comprehensive algorithm. This index information plays a critical role in determining whether feature points contribute to the creation of the global 3D point cloud map. In this process, we compare the previously



	mIoU	Params	FLOPs	FPS
SegFormer-B0	**37.4**	**3.7M**	**8.4G**	**50.5**
FCN-R50	36.1	49.6M	198.0G	23.5
SegFormer-B2	**46.5**	**27.5M**	**62.4G**	**24.5**
DeeplabV3+/R101	44.1	62.7M	255.1G	14.1
HRNet-W48 + OCR	43.0	70.5M	164.8G	17.0
SegFormer-B4	**50.3**	**64.1M**	**95.7G**	**15.4**
SETR	48.6	318.3M	362.1G	5.4

Fig. 5. Overview and comparison of the performance and processing speed of different neural networks for semantic image segmentation on the ADE20K dataset. The figure demonstrates the higher performance of SegFormer compared to DeepLabV3+. [20, p. 1]

assigned indices, keeping feature points with values of 0 and 1, while excluding those with a value of -1 from further consideration.

$$FeaturePointWithClasses = \begin{cases} Deleted, \text{ for Value } \textit{-1} \\ Keeped, \text{ for Value } \textit{0} \text{ and Value1} \end{cases}$$

As ORB-SLAM3 progresses, feature points that become map points are assigned additional labels such as "road" or "no road" depending on the values assigned. In the final phase of our current implementation, we need to evaluate the impact of semantic segmentation on the scaling of the global 3D point cloud map and the accuracy of the trajectory. Figure 6 gives a first qualitative insight into the results of our pipeline.

5 Evaluation

Environment mapping is done using different stereo cameras such as the *ZED2i* equipped with a polar filter and a baseline of 120 mm, and the *Intel RealSense D455* with a baseline of 95 mm. ORB-SLAM3 is applied to these cameras. The results presented in the images were obtained using the Intel Realsense D455. Detected features are recorded and stored in a three-dimensional point cloud. The impact of the low texture of the environment must be taken into consideration. To overcome the repetition of scene images, traffic cones are placed along the test track with gaps of 20 m, alternating in number on either side of the track. This ensures that feature matching and proper orientation in the point cloud occur. The camera is mounted at the rear-view mirror height on the top of the windshield, as is typical for cars. Figure 7 displays the test setup and track.

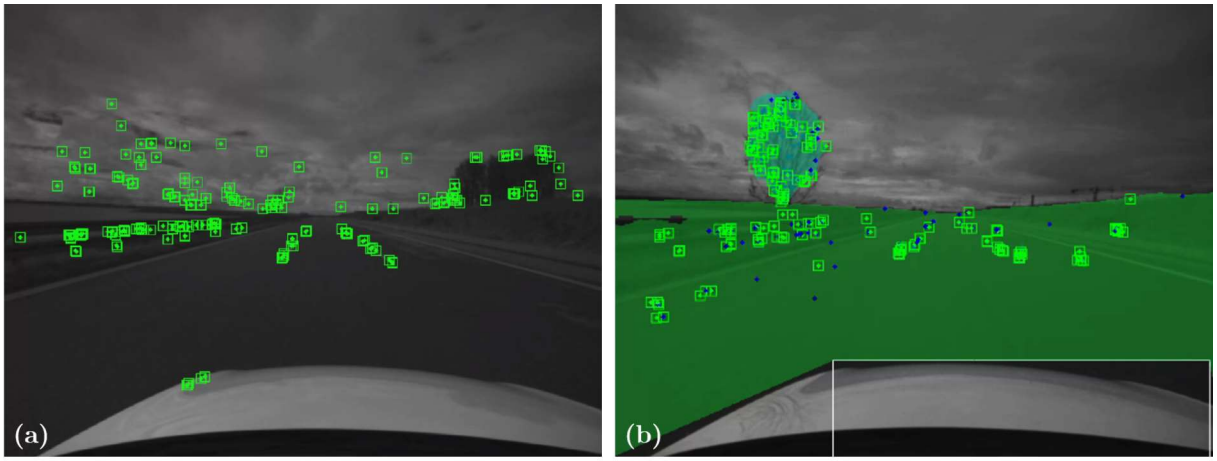


Fig. 6. ORB-SLAM3 detection results without filtering (a) and with filtered feature points (green rectangles) based on the semantic of the image (b). Using this method the false detection feature points in the sky and on the hood of the car are removed and being ignored by the mapping algorithm leading to better scaling and relocalization.

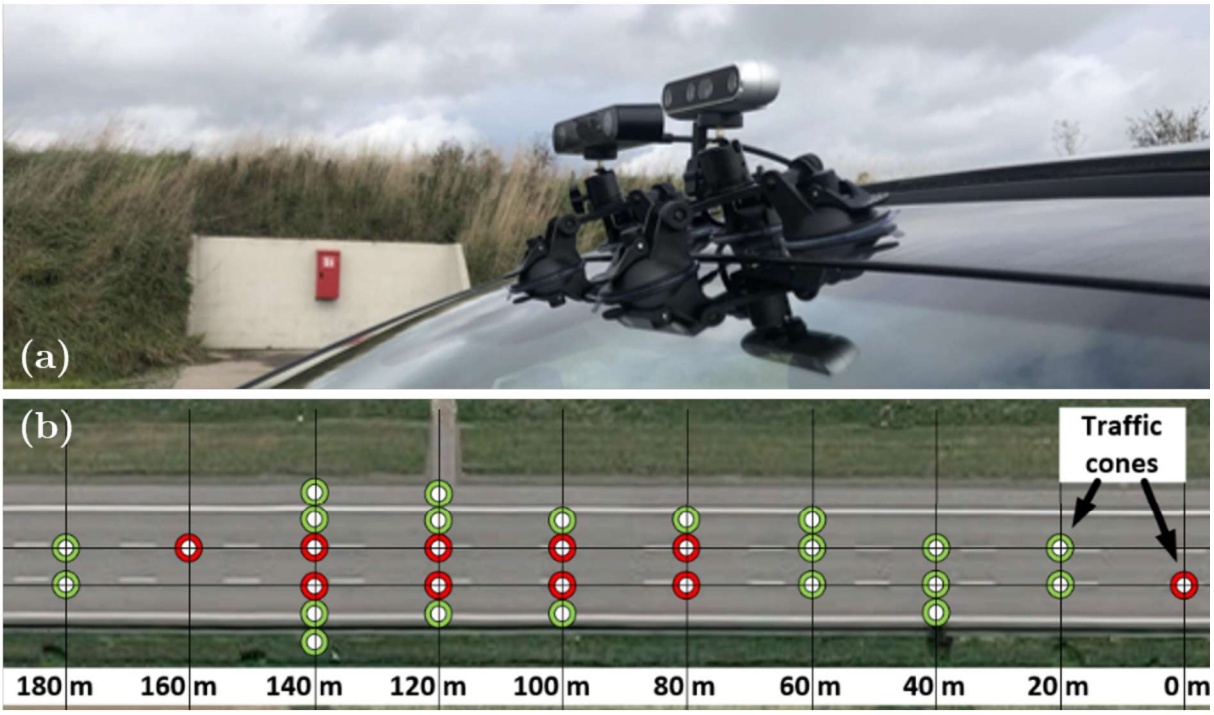


Fig. 7. Mounted stereo cameras on the test vehicle (a) and the test track with equipped traffic cones (b). The intervals are used in the evaluation to test the precision regarding scaling of the SLAM algorithm.

First of all, it will be examined whether semantic segmentation has a significant influence on the number of feature points. Table 1 shows, that the number of detected feature points decreased by more than 50 %. Assuming that the semantic image segmentation is correct, this means that more than 50 % of the detected feature points are located in the sky or on the hood of the vehicle and are therefore not suitable for further processing.

Table 1. Comparison of the number of detected map points with and without semantic segmentation.

Dataset	**Number of Detected Map Points**
Without Semantic Segmentation	8361
With Semantic Segmentation	3990
Total Deviations	-4371
Deviations in %	-52,3

To evaluate the impact on scaling, the distances between traffic cones are shown in table 2. The beneficial effects of semantic segmentation are easily seen. For an interval of 20 m, the average error is only -0.129 m compared to 1.414 m without segmentation. This leads to an relative reduction of -0.642 % compared to -7.073 % without segmentation. For the overall distance of the test track of 160 meters with semantic segmentation, the difference is -1.027 m (without segmentation 11.117 m). In summary, the use of semantic segmentation clearly adds significant value. Nevertheless, some additional considerations to the proposed approach are presented in the following section.

Table 2. Comparison of ground truth data (GT) with the original ORB-SLAM3 (No Segmentation) and with our approach (Segmentation). The intervals are given by the distances between the traffic cones as shown in Figure 7.

Interval	**GT Length**	**No Segmentation**	**Segmentation**
20 m - 40 m	20 m	18.603 m	19.840 m
40 m - 60 m	20 m	18.769 m	19.631 m
60 m - 80 m	20 m	18.974 m	19.921 m
80 m - 100 m	20 m	18.871 m	19.924 m
100 m - 120 m	20 m	18.871 m	19.993 m
120 m - 140 m	20 m	18.228 m	19.992 m
140 m - 160 m	20 m	18.365 m	19.678 m
160 m - 180 m	20 m	17.967 m	19.994 m
Average		18.586 m	19.871 m
Total Deviations		-1.414 m	-0.129 m
Deviations in %		-7.073	-0.642

6 Conclusion and Future Work

The proposed solution to increase the robustness and precision of vSLAM in automotive systems proved to be effective. Thus, the scaling error of the map could be significantly reduced by removing false-positive feature points. Furthermore, the information content of the map could be further increased by using semantic segmentation. The classes *Road* and *No Road* are now assigned to the corresponding feature points, which can be used in further calculations. Optimization opportunities arise with regard to processing speed, and image matching effects in homogeneous test scenarios as well as in more complex test procedures, which will be incorporated into Euro NCAP test scenarios in the future. Two proposals are being developed to improve the mapping of the environment:

- Change ORB-SLAM's feature detector and feature descriptor by a neuronal network to make the vSLAM more robust to a homogeneous environment as well as to rotations.
- Increase the processing speed of the entire vSLAM approach.

Acknowledgements

This work was supported with financial founds of the German Federal Ministry of Economic Affairs and Climate Action (BMWK) and of the project "KIIWI" (reference number: 16DHBKI060) which is funded by the German Ministry of Education and Research (BMBF). We also thank our partners EVOMOTIV GmbH, an automotive company in Germany.

References

1. Pendleton, S.D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y.H., Rus, D., Ang, M.H.: Perception, Planning, Control, and Coordination for Autonomous Vehicles. Machines **5**(1) (2017)
2. Weber, M., Weiß, T., Gechter, F., Kriesten, R.: Use of Artificial Intelligence and Image Segmentation for 3-Dimensional Modeling. In The Upper Rhine Artificial Intelligence Symposium UR-AI 2021, ed.: Artificial Intellgence – Application in Life Sciences and Beyond. (2021) 212–223
3. Weber, M., Weiß, T.: Object Detection by Using Image Segmentation for Augmented Reality in Advanced Driver Assistance Systems. In Institute of Energy Efficient Mobility, ed.: Reports on Energy Efficient Mobility – Volume 3. Volume 3. (2023) 14–28
4. Gao, X., Zhang, T.: Introduction to Visual SLAM: From Theory to Practice. 1st ed. 2021 edn. Springer eBook Collection. Springer Singapore and Imprint Springer, Singapore (2021)
5. Taketomi, T., Uchiyama, H., Ikeda, S.: Visual SLAM algorithms: a survey from 2010 to 2016. IPSJ Transactions on Computer Vision and Applications (2017)
6. Cheng, J., Zhang, L., Chen, Q., Hu, X., Cai, J.: A review of visual SLAM methods for autonomous driving vehicles. Engineering Applications of Artificial Intelligence **114** (2022) 104992
7. Campos, C., Elvira, R., G\´omez, J.J., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. IEEE Transactions on Robotics **37**(6) (2021) 1874–1890
8. Mur-Artal, R., Montiel, J., Tardos, J.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics (2015) 1147–1163
9. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7) (2022) 3523–3542

10. Wang, L., Shi, J., Song, G., Shen, I.f.: Object Detection Combining Recognition and Segmentation. In Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H., eds.: Computer Vision – ACCV 2007, Berlin, Heidelberg, Springer Berlin Heidelberg (2007) 189–199
11. Kuruvilla, J., Sukumaran, D., Sankar, A., Joy, S.P.: A review on image processing and image segmentation. In R., R., ed.: Proceedings of 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), Piscataway, NJ, IEEE (2016) 198–203
12. Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P.: Panoptic Segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, IEEE (2019) 9396–9405
13. van Ratingen, M.R.: The Euro NCAP Safety Rating. In Piskun, A., ed.: Karosseriebautage Hamburg 2017, Wiesbaden, Springer Fachmedien Wiesbaden (2017) 11–20
14. Yu, C., Liu, Z., Liu, X.J., Xie, F., Yang, Y., Wei, Q., Fei, Q.: DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments. In: Towards a robotic society, Piscataway, NJ, IEEE (2018) 1168–1174
15. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., eds.: Computer Vision – ECCV 2018, Cham, Springer International Publishing (2018) 833–851
16. Kamann, C., Rother, C.: Benchmarking the Robustness of Semantic Segmentation Models. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 8825–8835
17. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, IEEE Computer Society (2017) 1800–1807
18. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene Parsing through ADE20K Dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, IEEE Computer Society (2017) 5122–5130
19. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic Understanding of Scenes Through the ADE20K Dataset. International Journal of Computer Vision **127**(3) (2019) 302–321
20. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Neural Information Processing Systems (NeurIPS). (2021)
21. NVIDIA Developer: Robust Perception with Vision Transformer SegFormer (2023)