The ADUULM-Dataset - A Semantic Segmentation Dataset for Sensor Fusion

Andreas Pfeuffer
andreas.pfeuffer@uni-ulm.de
Markus Schön
markus.schoen@uni-ulm.de
Carsten Ditzel
carsten.ditzel@uni-ulm.de
Klaus Dietmayer
klaus.dietmayer@uni-ulm.de

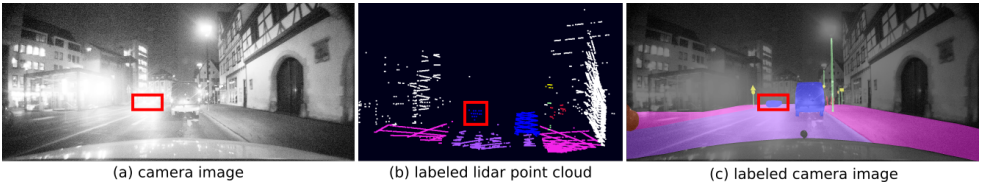
Institute of Measurement, Control and Microtechnology
University of Ulm
Ulm, Germany

Abstract

One of the key challenges of today's semantic segmentation approaches is to obtain robust and reliable segmentation results not only in good weather conditions, but also in adverse weather conditions such as darkness, fog or heavy rain. For this purpose, multiple sensor data of several sensor types such as camera and lidar are required to compensate the weather sensitivity of individual sensors. Hence, a semantic segmentation dataset is necessary, which contains camera and lidar data, but until recently, no such dataset exists. Therefore, the ADUULM dataset was created, a semantic segmentation dataset which consists of fine-annotated camera data and pixel-wise labeled lidar data recorded in diverse weather conditions. Additionally, the corresponding GPS, IMU and stereo information are provided, and for each annotated data sample, a short video-sequence is available, too. Furthermore, state-of-the-art semantic segmentation and drivable area detection approaches are evaluated on the proposed dataset, and it turned out that new methods are required to obtain robust and reliable results in adverse weather conditions. The ADUULM-dataset will be available online at https://www.uni-ulm.de/in/iui-drive-u/projekte/aduulm-dataset/.

1 Introduction

In recent years, several public available semantic segmentation datasets for autonomous driving have been released. For example, the Cityscapes dataset [8] published in 2016 is one of the first large datasets for semantic segmentation containing about 5000 fine-annotated camera images and the corresponding depth value. The data were recorded in good weather conditions in several cities in Germany. Recent published datasets such as Berkeley Deep Drive (BDD) [21], Apollo Scapes [10], and the Mapillary dataset [13] consist of even more labeled data and are also recorded in different weather conditions such as daylight, rain and night, and in different environments such as city or countryside. An overview of the most common segmentation datasets is shown in Table 1. Each of the listed datasets delivers



(a) camera image (b) labeled lidar point cloud (c) labeled camera image

Figure 1: The oncoming car (red boxes) can be hardly recognized in the camera image (a) due to the darkness and the blending car headlights, while it can be clearly seen in the lidar data (b). The corresponding fine-annotated camera image is illustrated in (c)

Table 1: Public available semantic segmentation datasets for autonomous driving

dataset	images	camera	lidar	stereo	video	weather conditions
CamVid [7]	701	✓	×	×	✓	daylight
Cityscapes [6]	5000	✓	×	✓	✓	daylight
Kitti [8]	400	✓	(✓)*	(✓)*	×	daylight
Mapillary [13]	25000	✓	×	×	×	diverse weather
BDD [21]	5683	✓	×	×	✓	diverse weather
Apollo Scapes [10]	55884	✓	×	✓	✓	diverse weather
SemanticKITTI [2]	43552	×	✓	×	✓	daylight
A2D2 [9]	41272	✓	✓	×	×	daylight
proposed	3893	✓	✓	✓	✓	diverse weather

*for 200 images, lidar and stereo data are available through the stereo benchmark.

pixel-wise annotated labels of the camera images and many of them also provides video sequences and stereo information. These datasets also contribute to the increasing success of the segmentation approaches such as ICNet [23], DeepLabv3+ [3] and BiSeNet [20], which perform very well in good weather conditions. Although some of the datasets contain adverse weather data, most of the semantic segmentation approaches are only suitable for a limited degree in adverse weather conditions such as snow, fog, rain or darkness. Therefore, a robust and reliable semantic segmentation of the car's surrounding is not possible [15], since the environment cannot be observed completely any more by camera systems due to sensor disturbances. For instance, in Fig. 1 (a), the van in front of the ego-vehicle can be clearly seen by the human eye, while the oncoming car can be hardly recognized due to the darkness and the blending car headlights. In contrast, the oncoming car can be clearly detected in the lidar data (see Fig. 1 (b)). Consequently, an appropriate sensor fusion of camera and lidar data might increase the segmentation performance and its robustness. Though fusing multiple sensor data of different sensor types is very popular in object detection tasks [4, 11, 14], there are hardly any similar methods for semantic segmentation applications. The reason is that until now there are no suitable datasets available (see Table 1), which contain camera and lidar data, except the Audi Autonomous Driving Dataset (A2D2) [9], which was published in the end of April 2020. This motivates us to construct the ADUULM (Autonomous Driving at University Ulm) dataset, a dataset, which consists of 3893 fine-annotated camera and lidar data and the corresponding GPS, IMU and stereo information. In contrast to the A2D2-dataset, the ADDULM-dataset also provides small video-sequences for each annotated data sample since the use of temporal information increases the segmentation accuracy further [16, 17, 19, 22]. Furthermore, the ADDULM-dataset was recorded in diverse weather condi-

tions such as daylight, darkness, (heavy) rain and fog, while the A2D2-dataset only contains daylight recordings. In this work, more details about the proposed dataset is given and state-of-the-art approaches of two computer vision sub-tasks (semantic segmentation and drivable area detection) are evaluated on this new dataset. Nevertheless, this dataset is also suitable for other applications such as object detection, instance and panoptic segmentation due its labeling policy.

2 Dataset

2.1 Sensor Equipment

The data of this dataset were collected by our test vehicle, which contains current state-of-the-art sensors of the autonomous driving domain. Our car is equipped with four cameras, four lidar sensors, an IMU accelerations measurement unit and a high-precision GPS sensor. Three cameras (a wideangle camera and a stereo camera pair) of resolution 1920×1080 are mounted behind the windshield, and the fourth camera of resolution 1392×1040 is mounted in the rear window. Each of the cameras provides 12 bit grayscale images at a frame rate of $15Hz$ and is triggered by the GPS time. The camera images are stored as 16bit images due to processing reasons. Moreover, the stereo information is yielded by the hardware unit ZYNQ Box SCS FPGA V. 1.0, in which the stereo global matching (SGM) approach is efficiently implemented. Furthermore, four lidar sensors of type Velodyne VLP-16 and frequency $10Hz$ are mounted on four different positions on the roof of the test vehicle (front, left, right, and rear) to capture accurate 3D information of the car's surrounding. In recent recordings, the front and the rear lidar sensors were replaced by Velodynes VLP-32 to get more precise spatial environment information. For each 3D point of the point cloud, the corresponding intensity value is also provided. Our test vehicle also contains an IMU measurement unit providing the vehicle's velocity and acceleration at a frame rate of $50Hz$, and receives the current GPS position for localization.

2.2 Sensor Calibration

An accurate sensor calibration is a very important requirement for fusing different sensor data to obtain a good and robust monitoring of the environment. The provided sensors of this dataset are mounted at different locations in the car and deliver their data at different time and with different frequency. Therefore, the sensor data has to be calibrated spatially and temporally. Generally, all sensors are spatially calibrated to a global reference point, the origin of our vehicle coordinate frame, which is the center of the vehicle's rear axis. For instance, the transformation matrix from the image coordinate frame to the vehicle frame is determined by means of extrinsic camera calibration. For this purpose, we distribute several chessboards patterns of known size in the camera's field of view and measure the position of their inner corner points in respect to the vehicle frame by means of a laser rangefinder. The rotation and the translation are yielded by minimizing the average reprojection error of the determined 3D corner points to the corresponding image points, provided that the intrinsic camera parameters are known. Moreover, the lidar sensors are calibrated to each other by point cloud matching recorded at the same scene and to the vehicle coordinate frame by matching the point clouds at different positions given the precise GPS-position.

After the spatial calibration, the sensors have to be calibrated temporally, since the sensor

Table 2: Class Definition

class	id	description
car	2	cars, vans, mini-buses including driver
truck	3	incl. agricultural and construction machines
bus	4	buses and trams
motorbike	5	motorbike and mopeds including motorcyclist
pedestrian	6	with bags and luggage
bicyclist	7	bike and bicyclist
traffic-sign	8	traffic signs without poles
traffic-light	9	traffic lights without poles
road	10	incl. road marking and public parking spaces
sidewalk	12	incl. curb
pole	15	e.g. pole of traffic-signs, lightings,...
unlabeled	0	not labeled areas and everything else

recording time of the different sensors are distinct. In our test vehicle, the GPS time is used as global reference time, to which every sensor is calibrated, yielding a unique time stamp for each sensor data. Nevertheless, the frequency of the sensors is still different, e.g. the frequency of the camera is $15Hz$, while the frequency of the lidar sensor is $10Hz$. Consequently, the different sensor data are available at different time stamps. This problem can be solved by means of the so-called ego-motion compensation. The lidar points recorded at time t_v, are transformed by a transformation matrix T to the image recording time t_i. For the determination of T, the current vehicle speed v, the current vehicle acceleration a, the rotation angle ψ of the vehicle, and the time difference $\Delta t = t_i - t_v$ between the sensors are necessary. These vehicle specific data can be easily obtained from the IMU. According to [1], the transformation matrix T is calculated by:

$$T(\Delta t) = \begin{pmatrix} \cos(\dot{\psi}\Delta t) & -\sin(\dot{\psi}) & 0 & v\Delta t + \frac{1}{2}a\Delta t^2 - \frac{1}{6}v\dot{\psi}^2\Delta t^3 - \frac{1}{8}a\dot{\psi}^2\Delta t^4 \\ \sin(\dot{\psi}\Delta t) & \cos(\dot{\psi}) & 0 & \frac{1}{2}v\dot{\psi}\Delta t^2 + \frac{1}{3}a\dot{\psi}\Delta t^3 - \frac{1}{24}v\dot{\psi}^3\Delta t^4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (1)$$

Another problem is that the lidar points are not recorded at one time point, but during a time interval due to the 360°rotation of the sensor. Hence, each lidar point has a different time stamp. Therefore, the scanning range of lidar sensor is divided into N circle segments. For each circle segment, a separate transformation matrix is determined, and each point within the corresponding circle segment is transformed by this so that the lidar points are temporally calibrated to the camera images. For each lidar sensor, the uncalibrated and calibrated point clouds represented in the vehicle frame are available. Furthermore, the dataset also provides the transformation matrices of each sensor.

2.3 Classes and Annotation

The ADUULM-dataset provides fine annotations at pixel level for the wideangle camera image and pixel-wise labeled laser data of all four lidar sensors. Our 3893 annotated data

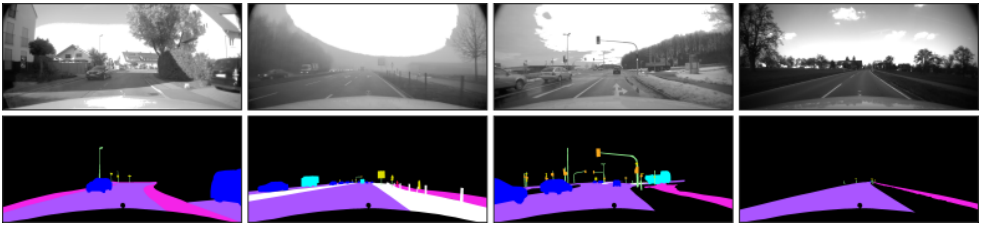


Figure 2: Examples of the ADDULM-dataset: wideangle camera image and the corresponding fine-annotated labels

samples are labeled by our in-house labeling tool and are double checked to guarantee highest quality. The 12 labeled classes were carefully selected according to their relevance in the autonomous driving application area and are defined in accordance to the Cityscapes dataset [5]. An overview of the provided classes can be found in Table 2. The camera images are annotated by manually drawing polygons onto the image, effectively capturing the contours of all unobstructed entities before assigning them an unambiguously class label. Every image is labeled from back to front so that a depth-ordering of the annotated objects is available, which is required for instance segmentation for example. In contrast, the lidar point clouds are annotated by means of a semi-automated procedure: the temporally synchronized three-dimensional laser point cloud are projected into the annotated camera image, i.e. into a two dimensional subspace, facilitating a first rough association of individual points with object classes. During this step, points located outside of the camera's field of view were neglected for the sake of computational efficiency. The misclassified points of the automated annotation are improved in a second step by a manually inspection of the scene by making use of various intuitive perspectives in the annotation tool. The class information is stored as fourth dimension of the point cloud, so that each annotated lidar point is represented by the tuple (x, y, z, id), where x, y, z are the coordinates of the laser point and id denotes the corresponding label id. Note, that only lidar points within the field of view of the wideangle camera and in a range of about $70m$ in front of the ego-vehicle are labeled, since a clear allocation is not possible for more distant object points due to their sparsity. Fig. 1 and Fig. 2 show some examples of the provided dataset.

2.4 Data Acquisition and Datasets

The data of this dataset were recorded from our test vehicle while driving in and around Ulm (Germany) at different daytimes, diverse seasons and in several good and adverse weather conditions such as fog, snow or (heavy) rain. Several data samples are periodically extracted from the recorded sequences at intervals from 5 seconds, which are then annotated as described in Section 2.3. Each data sample includes the camera images, the stereo information, the 3D point clouds of the four laser sensors, the current IMU data and the corresponding labels of the wideangle camera and of all four laser sensors. Furthermore, a short video sequence is provided for each data sample, which are stored as ROS-bags [13], so that the previous and following sensor information are available for a short period. All in all, the ADUULM dataset consists of 3893 fine-annotated data samples extracted from 100 video sequences. The 2012 good-weather samples are divided into two subsets for training and validation. The training set consists of 1504 samples (40 sequences) and the validation set

of 508 samples (15 sequences). We ensured that each of the subsets contain different representative scenarios such as urban and rural street scenes, which are recorded at different locations. Furthermore, there are 1881 adverse-weather data (45 sequences), which are used exclusively for testing. Training with the provided adverse weather data would increase the performance in similar situations, but sensor disturbances in adverse weather conditions are very complex so that it is almost impossible to capture all occurring weather scenarios. Hence, we are not interested in achieving high scores in special adverse weather scenarios. In contrast, our goal is to evaluate the robustness of current segmentation approaches, if the sensor data are disturbed by unknown noise as it occurs in adverse weather conditions. Therefore, the adverse weather data are excluded from the training and validation set in this dataset. As previously stated, the front and the rear lidar sensors were replaced by Velodynes VLP-32 to get more precise spatial environment information. In more detail, 410 of the 1504 training data, 117 of the 508 validation data and 1378 of the 1881 test data contain measurements of the Velodyne VLP-32, the remaining data were recorded by the original sensor setup.

2.5 Evaluation Metrics

For evaluation and comparison of semantic segmentation approaches, meaningful evaluation metrics are necessary, and hence, popular standard metrics are used for evaluation. A commonly used measure in this field is the percentage of correctly annotated image pixels or lidar points, which is known as pixelwise accuracy (acc.) and is given by

$$\text{acc} = \frac{\text{TP}}{\#\text{ pixels}} \tag{2}$$

where TP are the correctly classified image pixels or lidar points according to the ground-truth. A further popular evaluation metric is mean Intersection-over-Union (mIoU) [6], which is also known as mean Jaccard Index, and is more appropriate for datasets with imbalanced classes. The mIoU is defined as

$$\text{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{3}$$

where C is the number of classes, and TP_i, FP_i and FN_i are the true positive, false positive and false negative predictions for class i.

3 Semantic segmentation

In this section, different state-of-the-art semantic segmentation approaches are evaluated on the proposed dataset by means of the evaluation metrics pixelwise accuracy and mIoU (see Section 2.5). For this purpose, three popular image segmentation approaches are selected, which can be applied in real-time applications such as autonomous driving: DeepLabv3+ [3] (backbone ResNet-18), BiSeNet [20] (backbone Xception-39) and ICNet [23]. The ICNet is additionally trained on the provided lidar data, where a dense depth map was determined for each sample as described in [15], so that the benefit of the lidar data in some adverse weather scenarios can be demonstrated hereinafter. Furthermore, two recent published approaches are considered based on sensor fusion or video-segmentation, which are

Table 3: Evaluation of state-of-the-art segmentation approaches on the proposed ADUULM-Dataset

Approach	all		sunny		adverse	
	acc.	mIoU	acc.	mIoU	acc.	mIoU
BiSeNet [20]	92.63	42.69	**95.56**	50.16	79.60	22.32
DeepLabv3+ [3]	93.02	41.13	95.46	46.25	82.54	25.01
ICNet (img) [23]	92.51	45.63	95.17	51.57	79.52	23.43
ICNet (lidar) [23]	88.70	31.79	91.34	35.66	69.76	17.25
ICNet_LateFusion [15]	**94.07**	**50.56**	95.24	53.71	**87.17**	**31.51**
LSTM-ICNet [17]	93.14	49.60	95.29	**54.36**	85.22	31.13

Approach	night		rain		foggy	
	acc.	mIoU	acc.	mIoU	acc.	mIoU
BiSeNet [20]	57.37	6.28	83.35	26.23	94.38	40.34
DeepLabv3+ [3]	61.15	10.34	88.35	29.94	**95.69**	41.65
ICNet (img) [23]	56.81	8.00	83.34	29.35	94.63	42.15
ICNet (lidar) [23]	73.52	21.01	76.13	16.74	63.84	11.49
ICNet_LateFusion [15]	**78.06**	**24.80**	87.87	30.89	93.65	38.39
LSTM-ICNet [17]	69.55	17.84	**88.61**	**35.61**	95.28	**43.74**

more robust in adverse-weather conditions than the conventional segmentation methods: the ICNet_LateFusion [15], which is one of the first semantic segmentation approaches using camera and lidar data, and the LSTM-ICNet, which captures temporal image information by means of convLSTM-cells. All approaches were pretrained on Cityscapes, and then fine-tuned on the 1504 training data of the ADUULM-dataset for 100k iterations using a batch-size of two due to memory reasons. Each training was repeated for five times to compensate the training fluctuations. The remaining training hyper-parameters are identical to [15] and [17]. Table 3 shows the results of various image-sets evaluated in different weather conditions on the proposed dataset, and some qualitative examples are shown in Fig. 3. The image-set *all* contains good and adverse weather data, while the image-set *sunny* consists of only good weather data and the image-set *adverse* of adverse weather data. Furthermore, the performance in different adverse weather conditions (night, rain, foggy) is given. It turns out, that all segmentation approaches perform similarly in good-weather conditions. In contrast, their performance declines enormously in adverse weather conditions. For example, the single-image segmentation approaches ICNet, BiSeNet and DeepLabv3+ achieve an pixelwise accuracy of over 95% in good weather conditions, while only less than 80% of all pixels can be correctly classified in adverse weather conditions. The quantitative analysis also shows the use of multiple sensor-data and previous sensor information increase the segmentation accuracy in adverse weather conditions. For instance, ICNet_LateFusion outperforms the image based ICNet by about 20 percent in terms of pixelwise accuracy and by about 16 percent in terms of mIoU at night. Moreover, LSTM-ICNet increases the segmentation accuracy by about 6 percent in terms of pixelwise accuracy and by about 5 percent in terms of mIoU compared to the ICNet, if it rains. Nevertheless, both approaches - ICNet_LateFusion and LSTM-ICNet - are still not appropriate for a robust semantic segmentation in adverse weather conditions due to their poor performance, and hence, further research is necessary in the future.

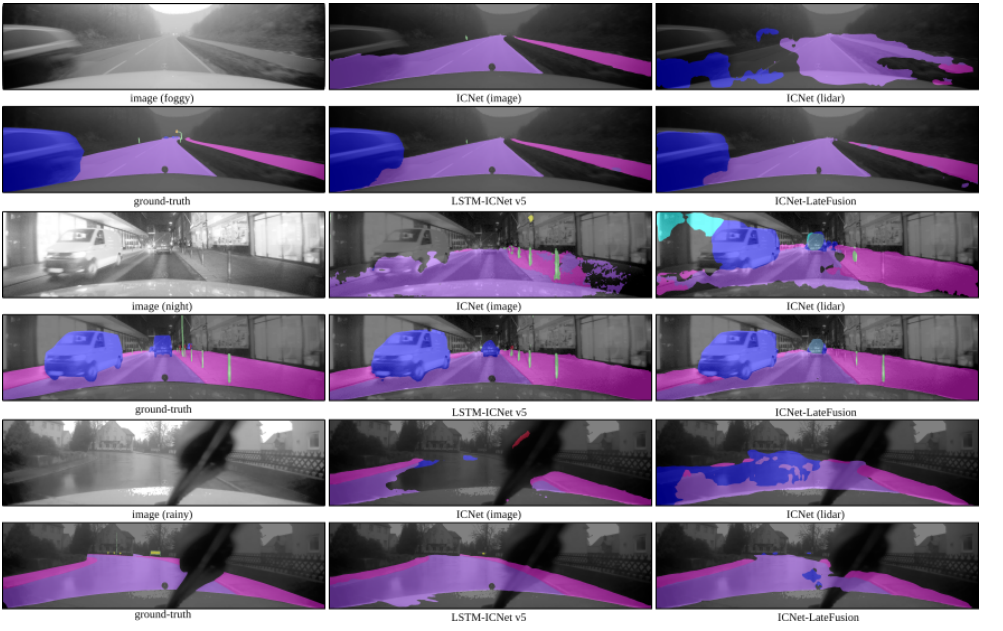


Figure 3: Qualitative results of state-of-the-art semantic segmentation approaches on the ADUUlm-dataset

4 Drivable Area Detection

The task of drivable area detection is to identify regions, where the vehicle can drive based on its current environment. Unlike road segmentation, the focus of drivable area detection is to detect separate driving lanes and classify them based on their driving direction. Drivable area annotations are missing in most semantic segmentation datasets, which usually only define a road class. To our knowledge, the only available dataset with drivable area annotations is the BDD-Dataset [21]. BDD defines two classes of drivable area, direct and alternative drivable area. Direct drivable area is the lane, the ego vehicle is currently driving on. Alternative drivable area are all lanes, where the ego vehicle can potentially drive on, pointing in the same driving direction as direct drivable area. Apart from the two classes defined by BDD, an autonomous vehicle may need more information about the lane geometry such as if the lane is a turn lane or has a different driving direction. The ADUULM dataset builds on this idea, annotating a total of six classes for the drivable area detection task. Ego lane is the lane, where the ego vehicle is currently driving on. A parallel lane is a lane, which has the same direction as the ego lane. Parallel lanes, which are clearly identified as a turn lane, e.g. because of a turn arrow, are labeled as a separate class called parallel turn lane. Opposite lane is every lane, which points in the opposite direction than the current ego lane. Following the parallel turn lane, opposite turn lanes are also annotated as a separate class. Lanes which are used for parking only are annotated as parking lane. To cope with the difficulty of annotating crossing areas, where multiple lanes overlap or cannot be distinguished clearly, a separate class is introduced as crossing area. Identifying the crossing area even without knowing the exact lane geometry can be valuable, as the autonomous vehicle can pay more attention in these situations. Fig. 4 shows multiple examples of annotated images for the drivable area task. The ADUULM dataset contains a total of 2314 annotated camera images for this task.

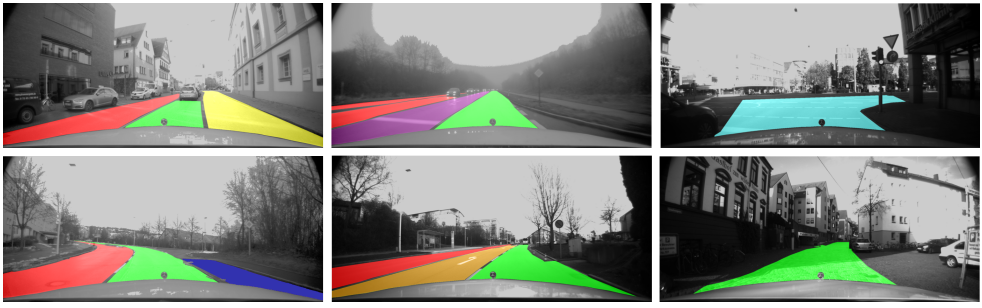


Figure 4: Annotation examples for the drivable area detection task. Ego lane is colored green, parallel lane yellow, parallel turn lane dark yellow, opposite lane red, opposite turn lane purple, parking lane blue and crossing area light blue.

Table 4: Evaluation of state-of-the-art segmentation approaches on the ADUULM drivable area detection task.

	default		simplified	
Approach	acc.	mIoU	acc.	mIoU
ICNet [23]	88.31	36.28	90.61	71.02
DFANet [12]	89.50	41.72	90.25	68.93

4.1 Task Definition

We define two drivable area detection tasks. The first contains all six classes described above and a background class, leading to a total of seven different classes. We refer to this task as the default drivable area detection task. Furthermore, we propose a simplified version with only four classes in total. The classes ego lane and crossing area are joint to a new ego lane class. Pixels of the classes parallel turn lane and opposite turn lane are considered part of the classes parallel lane and opposite lane respectively. Parking lane pixel are considered part of the background class. We refer to this task as the simplified drivable area detection task. Since the drivable area detection task is a segmentation task, we use the same metrics as for the semantic segmentation task, accuracy and mIoU. Refer to Section 2.5 for details.

4.2 State-of-the-art Results

In this section, we evaluate two different real-time state-of-the-art architectures on the drivable area task, ICNet [23] and DFANet [12]. We split the dataset in train, val and test set and report performance on the test set. We use 1545 images for the train set, 569 images for the val set and 200 images for the test set. The training protocols follow the settings in [23] and [12]. Results are shown in Table 4 for the default and simplified detection task. Both networks have poor performance on the default task. This states that state-of-the-art segmentation networks cannot solve the task properly, which leaves space for future research in this area. Qualitative results on the simplified task are shown in Fig. 5. The top row shows the prediction, while the bottom row shows the ground truth. The network can extract the general lane geometry but struggles at further distances and in separating different lanes of the same class.

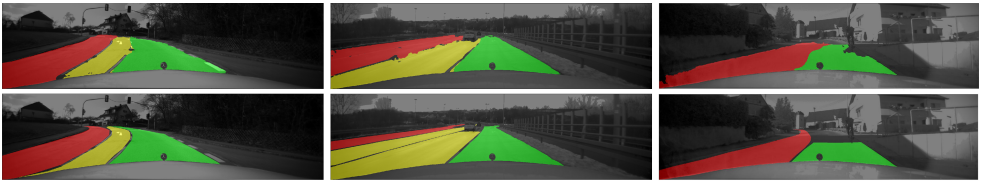


Figure 5: Qualitative results of DFANet on the simplified drivable area detection task. The top row shows the prediction while the bottom row shows the ground truth. The mid and right columns show more challenging examples.

5 Conclusion

In this work, the ADUULM-dataset was introduced, one of the first semantic segmentation datasets, which contains fine-annotated camera images and point-wise labeled lidar data and additionally provides video-data. Several state-of-the-art approaches in the field of semantic segmentation and drivable area detection were evaluated on the proposed dataset, but generally, the dataset can also be used for diverse other tasks such as object-detection or panoptic segmentation. The experiments show that the considered approaches perform well in good weather conditions but fail in adverse weather conditions. Therefore, new approaches are necessary to improve the segmentation approaches, e.g. advanced sensor fusion and video segmentation approaches. Finally, we hope the proposed dataset contributes to making the segmentation approaches more robust against unknown sensor disturbances and adverse weather effects so that a reliable segmentation will be possible in all weather conditions in the future.

References

[1] Alexander Barth. *Vehicle tracking and motion estimation based on stereo vision sequences*. PhD thesis, Rheinischen Friedrich-Wilhelms-Universität zu Bonn, 2010.

[2] 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.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. *CoRR*, abs/1611.07759, 2016. URL http://arxiv.org/abs/1611.07759.

[5] 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. *CoRR*, abs/1604.01685, 2016. URL http://arxiv.org/abs/1604.01685.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

[7] Julien Fauqueur, Gabriel Brostow, and Roberto Cipolla. Assisted video object labeling by joint tracking of regions and keypoints. In *IEEE International Conference on Computer Vision (ICCV'2007)*, 2007.

[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] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. 2020. URL https://www.a2d2.audi.

[10] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. *arXiv: 1803.06184*, 2018.

[11] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Lake Waslander. Joint 3d proposal generation and object detection from view aggregation. *CoRR*, abs/1712.02294, 2017. URL http://arxiv.org/abs/1712.02294.

[12] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2019.

[13] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. URL https://www.mapillary.com/dataset/vistas.

[14] Andreas Pfeuffer and Klaus Dietmayer. Optimal sensor data fusion architecture for object detection in adverse weather conditions. In *2018 21st International Conference on Information Fusion (FUSION) (FUSION 2018)*, pages 2592–2599, Cambridge, United Kingdom (Great Britain), July 2018.

[15] Andreas Pfeuffer and Klaus Dietmayer. Robust semantic segmentation in adverse weather conditions by means of sensor data fusion. In *2019 22nd International Conference on Information Fusion (FUSION) (FUSION 2019)*, Ottawa, Canada, July 2019.

[16] Andreas Pfeuffer and Klaus Dietmayer. Robust semantic segmentation in adverse weather conditions by means of fast video-sequence segmentation. *2020 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2020.

[17] Andreas Pfeuffer, Karina Schulz, and Klaus Dietmayer. Semantic segmentation of video sequences with convolutional lstms. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1253 – 1259, 2019.

[18] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Ng. Ros: an open-source robot operating system. volume 3, 01 2009.

[19] Sepehr Valipour, Mennatullah Siam, Martin Jägersand, and Nilanjan Ray. Recurrent fully convolutional networks for video segmentation. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 29–36, 2017.

[20] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *CoRR*, abs/1808.00897, 2018. URL http://arxiv.org/abs/1808.00897.

[21] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018. URL http://arxiv.org/abs/1805.04687.

[22] E. E. Yurdakul and Y. Yemez. Semantic segmentation of rgbd videos with recurrent fully convolutional neural networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 367–374, Oct 2017. doi: 10.1109/ICCVW.2017.51.

[23] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. *CoRR*, abs/1704.08545, 2017. URL http://arxiv.org/abs/1704.08545.