2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Early Fusion of Camera and Lidar for robust road detection based on U-Net FCN

Florian Wulff, Bernd Schäufele, Oliver Sawade
Fraunhofer Institute for Open
Communication Systems (FOKUS)
Berlin, Germany
{florian.wulff, bernd.schaeufele, oliver.sawade}@fokus.fraunhofer.de

Daniel Becker, Birgit Henke, Ilja Radusch
Daimler Center for Automotive
Information Technology Innovations (DCAITI)
Berlin, Germany
{daniel.becker, birgit.henke, ilja.radusch}@dcaiti.com

***Abstract*—Automated vehicles rely on the accurate and robust detection of the drivable area, often classified into free space, road area and lane information. Most current approaches use monocular or stereo cameras to detect these. However, LiDAR sensors are becoming more common and offer unique properties for road area detection such as precision and robustness to weather conditions. We therefore propose two approaches for a pixel-wise semantic binary segmentation of the road area based on a modified U-Net Fully Convolutional Network (FCN) architecture. The first approach UView-Cam employs a single camera image, whereas the second approach UGrid-Fused incorporates a early fusion of LiDAR and camera data into a multi-dimensional occupation grid representation as FCN input. The fusion of camera and LiDAR allows for efficient and robust leverage of individual sensor properties in a single FCN. For the training of UView-Cam, multiple publicly available datasets of street environments are used, while the UGrid-Fused is trained with the KITTI dataset. In the KITTI Road/Lane Detection benchmark, the proposed networks reach a MaxF score of 94.23% and 93.81% respectively. Both approaches achieve real-time performance with a detection rate of about 10 Hz.**

I. Introduction

A self-driving car has to be capable of sensing and interpreting its environment as well as navigating without any human input or control. This implies a need for a highly accurate perception and understanding of the environment under any circumstances. Despite the existence of advanced driver assistance systems (ADAS) algorithms for specific scenarios (e.g. highways), autonomous driving in unknown, complex and highly dynamic urban scenarios is still a big challenge. Currently automated urban driving is undergoing first real-world tests. One of the basic building blocks for subsequent tasks like mapping, path planning and control is scene understanding and a model of the environment. This includes accurate and reliable road and lane detection as well as reliable assessment of the potentially drivable collision-free area. Especially for automation levels 4 and 5 [1], where no human driver is monitoring the environment and system operation, a correct estimation of the drivable area has to be made in any situation.

We differentiate between three perception levels: Ego lane, road area and free space. The ego lane describes the current lane of the vehicle (if on-road), road area is the sum of all available lanes (including oncoming traffic) and free space is the potentially drivable flat ground area. All three levels are relevant to achieve a robust perception system and can serve as fallback solutions for each other. The vehicle uses ego lane information while in lane keeping and the road area information for selection of other modes (e.g. lane changes, overtaking). For free-space detection, we utilize a simple complimentary approach to serve as sanity check and fallback. Figure 1 illustrates the predicted area of those three levels.

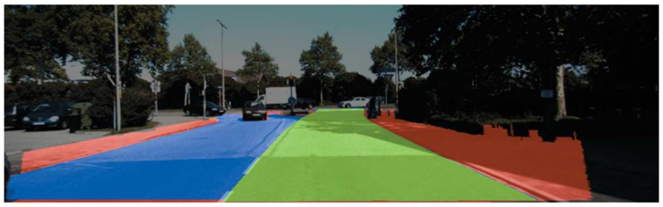

Fig. 1: Overview of levels: red shows free space, blue shows road area and green shows ego lane (overlaid in this order).

Camera sensors have become prevalent in road area and lane detection as they offer highly detailed information and are well suited to evaluation with classical computer vision methods as well as with Fully Convolutional Networks (FCNs). Thus, large training sets exist. The LiDAR sensor in comparison is very sparse in resolution yet it also has exceptional accuracy and more robustness to environmental conditions.

In this work, we propose a FCN based on the U-Net [2] architecture for detection of the road area. To this end, we present two main configurations: UView-Cam uses the input from a single camera in the windshield viewpoint. In contrast, UGrid-Fused employs a fusion of LiDAR and camera data into a multi-dimensional occupation grid representation. While a multitude of training data is available for UView-Cam, only the KITTI road dataset [3] can be used for UGrid-Fused. The detection performance of both networks is evaluated in the KITTI Road/Lane Detection benchmark [4].

This paper is structured as follows: In Section 2, related work is introduced and compared to the proposed approach. In Section 3, the architecture, FCN training and optimization process of the UView-Cam approach is discussed. Section 4 explains the occupation grid data representation for the

This work was supported by the German Federal Ministry of Transport and Digital Infrastructure (BMVI).

UGrid-Fused approach and the adaptions to the FCN architecture. In Section 5, the performance for both approaches is discussed. We conclude the paper in section 6.

II. Related Work

Road detection is a highly active field of research and rapid progress is achieved, based currently on the advances of FCNs. In order to be able to quantitatively compare the performance of our different approaches, we use the open computer vision KITTI Road/Lane Detection benchmark [4]. In the following, we compare our approach against a number of the top submissions in the leaderboard [5] of this benchmark.

Several approaches exists that are solely based on cameras: DDN [6] is a 7-layered deep network for scene parsing of road pixels. It combines CNNs as feedforward layers for pooling with deep deconvolutional networks as feedback layers for unpooling. Up-Conv-Poly [7] is an efficient deep neural network introducing a new mapping between classes and filters for the deconvolutional part of the network. MultiNet [8] incorporates classification, detection and semantic segmentation simultaneously. TuSimple [9] uses hybrid dilation convolution (HDC) and the dense upsampling convolution (DUC).

Other related approaches are focusing on using only LiDAR data: In [10], road detection using occupancy grid mapping is proposed, with only 3D LiDAR data under various challenging environments. Feature maps are calculated as occupancy grids, which are including gradient values or height value averages. This allows approaching the road detection as a classification problem of different features using Markov-Random-Fields (MRF). LoDNN [11] is a network designed to perform the road segmentation task using only LiDAR data with a deep learning approach. A multi-dimensional occupancy grid with 6 layers encoding different statistics such as mean elevation and density of the reflections is extracted from a LiDAR scan and used as input for a FCN. In this work, we aim to apply state-of-the-art FCNs based on a combined camera and LiDAR occupancy grid representation to achieve a highly accurate and robust road detection performance.

III. Methodology UView-Cam

In the following, we describe the selected FCN architecture and the applied modifications, list the used data sets and explain the multi-faceted training and optimization process for the UView-Cam approach which uses a single camera view.

A. FCN architecture

To select a suitable architecture, we have tested and compared state-of-the-art FCNs in terms of their performance for a binary semantic segmentation task in camera images. In particular, we have examined FCN [12] and U-Net [2]: FCN is one of the most common and frequently applied networks for image segmentation tasks to date. [12] stated that FCNs perform best when using VGG [13] as encoder network. In contrast, U-Net performs well in binary segmentation with small datasets and has a relatively simple architecture and low number of parameters compared to FCN due to the lack of fully connected layers. Our proposed architecture is a combination of the VGG [13] and U-Net [2]. It adopts all convolutional layers from the VGG19 network as the encoder part. The decoder is equal to a mirrored encoder, using convolutions with the same kernels and shape after each upsampling step, as introduced in the U-Net. Finally, the concatenations between all associated layers in the encoder and decoder part are adapted from U-Net as well. The resulting network architecture is shown in Figure 2.

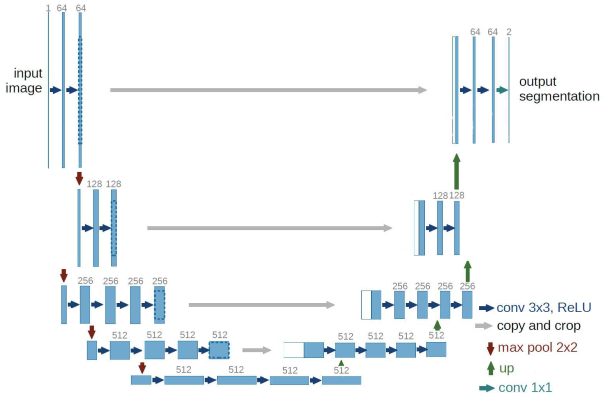

Fig. 2: Proposed VGG19-UNet architecture adapted from [2].

B. Datasets

In order to train the proposed FCN for the task of road detection, a large number of datasets with a good quality of pixel-wise labels is required. Thus, we have selected multiple publicly available datasets and converted the labels into a common binary representation (i.e. road, non-road) to differentiate the drivable road from the rest of the image.

The KITTI road detection evaluation [3] [4] provides annotations of 289 training and al 290 test images grouped into multiple categories. The dataset includes data from a monocular and a stereo vision camera, a monochromatic camera, a Velodyne HDL-64E LiDAR scanner and an IMU as well as calibrations between the different sensor views.

There are additional datasets for the task of the semantic segmentation: Hermans, A. and Floros, G. have annotated 203 images [14], Alvarez, J. has annotated 323 images [15] and Ros, G. has annotated 146 images [16]. Using the annotated RGB masks from these datasets, a binary mask for the single road class is extracted from each dataset. This results in a total of 672 additional images which are used for training. Some of the datasets contain lower quality ground truth masks or are labeled using different guidelines. For this reason, the extracted road masks have been checked and corrected manually to create a dataset of a high and consistent quality.

Also, the Cityscapes dataset [17] has been used which consists of annotations for 30 classes in 5000 fine annotated images and in 20000 coarse annotated images. As one of the 30 classes is the class ROAD and there are no other classes labeled on the street surface (e.g. lane markings, bicycle paths, etc.), the extraction of high-quality binary road masks is straightforward. Only the 5.000 fine images are used, as the coarse images are not labeled precisely enough to improve the overall quality of the whole dataset.

C. Optimizations

Due to the low amount of available training data, we have applied a systematic empirical optimization process to achieve the best possible performance. The applied optimization for the final classifier in the evaluation are as follows.

Firstly, to drastically reduce the computational effort necessary to train a network until convergence, pretrained networks weights from the Imagenet challenge [18] are used for all convolutional layers in the encoder part (cf. Figure 2). We trained the decoder in our network from scratch while encoder weights are simultaneously retrained. Pretrained weights are resulting in a faster convergence, which allows shorter training durations and a lower number of epochs as well as a higher overall accuracy. Secondly, we evaluated the choice of different color spaces. Using YUV has a beneficial effect because the color channels are not correlated with the contrast and the intensity, unlike in RGB. Also, YUV has no additional computational cost regarding the network performance and would also allow a training with grayscale images by using only the Y channel, therefore YUV is the best choice. Thirdly, we examined different possibilities for the normalization of input data. While it is optimal to normalize the stochastic properties to a mean of zero and a standard deviation of one over the total dataset, this cannot be applied to unknown test images or images from different datasets. Thus, the chosen normalization is done for each channel of every single image independently of the overall dataset. Fourthly, the chosen resolution of the input images has the largest effect of all changes on the accuracy of the network, but also significantly raises the inference time and training time. Resolutions like $1024px \times 256px$ or $512px \times 128px$ were evaluated in the training process to improve the inference time, but using small resolutions destroys significant amounts of the contained information in the image. Hence, for the final evaluation we use a resolution of $1248px \times 384px$ to reach optimal accuracy.

IV. Methodology UGrid-Fused

While a single camera can deliver excellent results for road detection under normal conditions, the quality and robustness of this approach can degrade quickly in the case of extreme environmental conditions, such as bright sunlight or snow. In these cases, it is beneficial to incorporate LiDAR scans. Thus, we introduce a multi-dimensional occupancy grid for early sensor fusion of camera and LiDAR data, which is used as input to the FCN. In the following, we describe the grid generation, adoption of the previously introduced U-Net architecture as well as training process.

A. Generation of LiDAR/camera fusion occupancy grids

An occupancy grid can be described as a grid M or array that models the local occupancy evidence of the environment by orthographically projecting the 3D sensor measurements on a 2D plane P, where P is oriented parallel to the road when assuming that the road surface is planar.

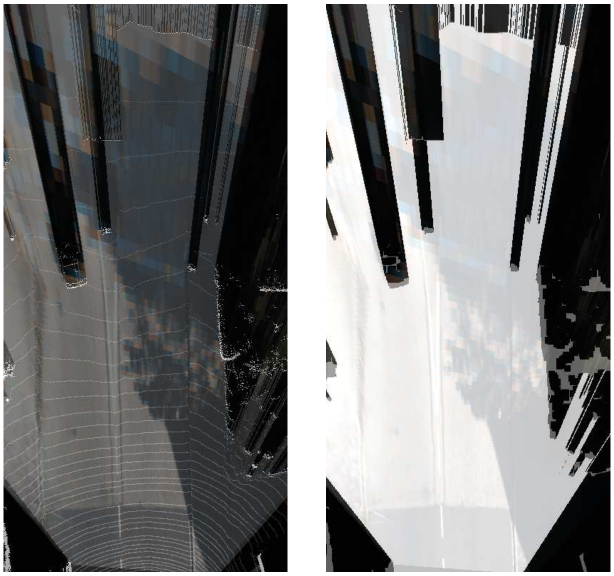

Fig. 3: Occupancy grid visualization in BEV, combined image and binary LiDAR scan (left), detected free space (right).

For each grid cell, 15 different statistics, based on the LiDAR scan points in each cell, are calculated. For a cell size of $5cm$, this yields an occupancy grid represented as matrix of the dimensions: $15 \times 800 \times 400$, which constitutes an area of $40m \times 20m$ in front of the car in the Bird-Eye View (BEV). The following occupancy grid maps are calculated:

1) Binary map: Each grid cell is either one if any reflection was measured for the coordinates of this cell in any layer or otherwise zero.
2) Count map: Each grid cell is between zero or one, depending on the accumulated sum of reflections in all layers divided by the number of total layers.
3) Obstacle map: Based on a minimum height threshold, each grid cell is either zero, if no height measurement is above the threshold or one otherwise.
4) Six height measurement maps: For each grid cell, local statistics based on the individual height measurements, are calculated (i.e. minimum, maximum, mean, minimum-maximum-difference, mean-standard-deviation, mean-variance).
5) Six reflectivity intensity maps: For each grid cell, local statistics based on measured LiDAR scan reflectivity/intensity are calculated (same statistics as for height measurement maps, see above).

This results in a total of 15 occupancy grids representing the LiDAR point cloud. The fusion of this grid with camera images, architectural changes and training procedures of the FCN are discussed in the following.

B. Sensor fusion

One of the main drawbacks of camera images is the 2D environment representation without depth information. To enhance the image and include additional information, the LiDAR point cloud can be added using the occupancy grid representation to the camera image. Conversely, we project the camera image into the BEV representation and overlay them onto the LiDAR occupancy grid. Thereby all measurements are in the same Cartesian coordinate system. Besides, this removes the dependency on the camera perspective and thus allows for augmenting the occupancy grid with additional sensor data or even map information (e.g. road lanes, landmarks) considering the current vehicle position. Moreover, algorithms can use the available data simultaneously from a single source with a unified metric representation.

When fusing camera images and LiDAR scans, one has to consider the different mounting positions, their point of views and the different coordinate systems of the sensors. By calibrating both sensors with regard to each other using the same scene or objects, a transformation matrix between the 3D polar coordinate systems of the LiDAR scan and the 2D Cartesian coordinate system of the camera image into a common 3D or 2D coordinate system can be calculated. This is achieved by applying an affine transformation based on matching real-world point pairs in both sensors. The resulting parameters are given in the KITTI Road dataset [4]. Figure 3 shows the combined camera image and occupancy grid.

C. FCN architecture and datasets

As training dataset, only the KITTI Road/Lane Detection dataset (cf. Section III-B) is used which contains a combination of 289 synchronized LiDAR scans and camera images. As described previously, the LiDAR and camera data is converted into an 18 channel occupancy grid in BEV which is used as input layer for the FCN. Also, ground truth class labels are available in the BEV.

The same VGG19-UNet network as for the UView-Cam approach (cf. Section III-A) is used. The only modifications in the network architecture are an adaption of the input size of the initial convolutional layer to the corresponding number of 18 image channels (i.e. 15 dimensions for the occupancy grids and a three channel color image) and different resolution for input and output layer. All other layers are unchanged.

D. Alternative usages

While this work focuses on road detection, the occupancy grid also allows for the previously introduced 3 perception levels (ego lane, road area, freespace, cf. Figure 1).

a) Freespace detection: A freespace map is a 2D representation of the flat and obstacle-free surface around the vehicle, which is considered drivable in this case. To extract the free space, the obstacle map is evaluated for the closest obstacle in each direction. Starting at the sensor position as the origin, for every angle in each direction, the distance to the first reflection is determined. All pixels on a straight line from the source till the obstacle border are considered unoccupied cells, the reflections by the obstacle are considered occupied cells and any pixels behind the first obstacle are considered unknown cells. The detected freespace is shown in Figure 3.

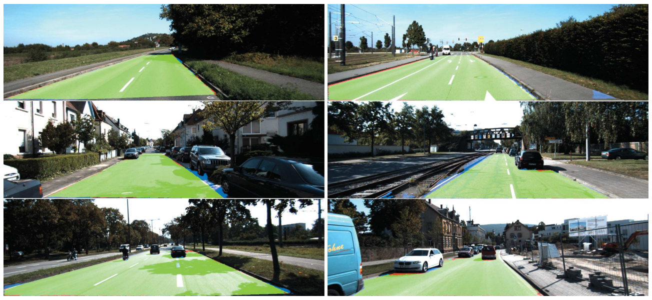

Fig. 4: Camera view visualization of test images: True Positives (green), False Negatives (red) & False Positives (blue).

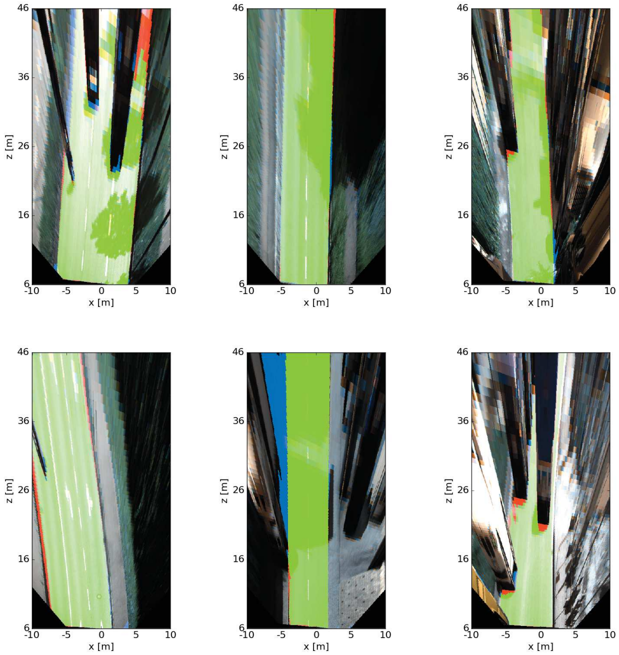

Fig. 5: BEV visualization of test images: True Positives (green), False Negatives (red) & False Positives (blue).

b) Ego-lane detection: The semantic segmentation using FCNs, as described in Section III, can not only be applied to the road segmentation but also for an end to end ego-lane segmentation. The KITTI Road/Lane detection challenge [5] provides 95 annotated training images and 96 test images for the lane detection task. Training is done using the VGG19-UNet network and the training parameter described in Section III-C for approximately 300 epochs. By using the already trained weights adapted from the road segmentations as pre-trained weights, a good performance is achieved despite the small number of images in the dataset.

V. Evaluation

In the following, we discuss the evaluation process and results of the two proposed road detection approaches. After explaining the experimental setup, we outline the applied

	MaxF	AP	PRE	REC	FPR	FNR	Inference time
TuSimple [9]	96.41%	93.88%	96.44%	96.37%	1.96%	3.63%	20ms
MultiNet [8]	94.88%	93.71%	94.84%	94.91%	2.85%	5.09%	17ms
UView-Cam (all datasets)	**94.23%**	87.98%	93.23%	95.24%	3.81%	4.76%	103ms
LoDNN [11]	94.07%	92.03%	92.81%	95.37%	4.07%	4.63%	18ms
Up-Conv-Poly [7]	93.83%	90.47%	94.00%	93.67%	3.29%	6.33%	80ms
UGrid-Fused (only KITTI dataset)	**93.81%**	89.49%	93.70%	93.91%	3.48%	6.09%	83ms
DDN [6]	93.43%	89.67%	95.09%	91.82%	2.61%	8.81%	2000ms
UView-Cam (only KITTI dataset)	**93.05%**	89.99%	93.66%	92.44%	3.45%	7.56%	103ms

TABLE I: KITTI urban road benchmark scores for our approaches (highlighted in bold) compared to related works from the leaderboard [5] (see also Section II), ranked by MaxF score.

evaluation metrics and process in accordance with the KITTI benchmark as well as the actual detection performance.

A. Experimental setup

The used hardware for implementation, training and evaluation of the proposed FCNs is a Core i7-2860QM @2.5GHz CPU, 16GB RAM, Crucial MX300 525GB Solid State Drive and NVIDIA TITAN X (Pascal) with 12 GB GDDR5X RAM graphics card. Also, the following software setup has been used: Ubuntu 16.04 64 bit Operating System, Python 3.5 programming language, gcc5.4.0 (c++14), OpenCV 3.1, Tensorflow 1.1 and Keras 2.0 libraries with CUDA 8.0 and CuDNN 6.0 hardware acceleration.

B. Evaluation metrics

As the road prediction is a binary classification, the evaluation of the performance can be done using standard metrics from the confusion matrix of two classes, the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The following metrics are defined to evaluate and compare the performance of the proposed approaches in accordance with the KITTI road detection benchmark [4].

$$\text{F-measure} = (1+\beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{1}$$

$$F_{max} = \underset{\tau}{\operatorname{argmax}} \text{ F-measure} \tag{2}$$

$$AP = \frac{1}{11} \cdot \sum_{r \in 0,0.1,\ldots 1} \max_{\tilde{r}:\tilde{r}>r} \text{Precision}(\tilde{r}) \tag{3}$$

$$PRE = \frac{TP}{TP+FP} \tag{4}$$

$$REC = \frac{TP}{TP+FN} \tag{5}$$

$$FPR = \frac{FP}{FP+TN} \tag{6}$$

$$FNR = \frac{FN}{TP+FN} \tag{7}$$

C. Evaluation process

The evaluation of the road detection performance is done in the BEV as explained in [4]. The BEV equally weights the predictions depending only on their real-world metric size anywhere in the image in contrast to perspective view, which would be heavily dependent on distance. Also, comparable evaluations between different sensors, resolutions and field of views are facilitated when transformed into a common BEV. Finally, the proposed road detection approaches were used to annotate the test dataset (i.e. 290 images) of the KITTI vision benchmark and subsequently upload the masks to the evaluation server. The resulting scores are shown in Table I. Also, a number of image visualizations illustrate the performance of the uploaded masks on the test dataset in the perspective view (see Figure 4) and the BEV (see Figure 5), which use a RGB scheme, where green, red and blue represent true positives, false negatives and false positives respectively.

D. Evaluation road detection results

Table I displays the results for our proposed approaches evaluated in the category URBAN_ROAD in comparison with the related work presented in Section II. Thereby, the *UView-Cam* approach (as presented in Section III) is operated in two configurations: A) trained with only the KITTI dataset and B) trained with all datasets (cf. Section III-B). Additionally, *UGrid-Fused* (as presented in Section IV) is only trained with KITTI data. Our proposed *UView-Cam (only KITTI dataset)* approach achieves a MaxF score of 93.05% and can be considered as baseline. Compared to this baseline, we achieve an improvement of 0.76% MaxF by also incorporating LiDAR data in the occupancy grid representation with the *UGrid-Fused (only KITTI dataset)* approach, reaching a MaxF score 93.81%. Ultimately, the *UView-Cam (all datasets)* approach, that is trained with multiple datasets, yields the best result with a MaxF score of 94.23%, i.e. a 1.27% improvement relative to the baseline approach. This benchmark demonstrates that our proposed road-detection approaches are achieving competitive results compared to other state-of-the-art approaches. While the *UGrid-Fused* approach slightly outperforms the *UView-Cam* approach based on the KITTI dataset, the best result is achieved by training the *UView-Cam* FCN with multiple datasets. Hence, we expect a significant improvement for the *UGrid-Fused* approach when trained with multiple datasets.

The KITTI benchmark is an invaluable tool for performance estimation, but is is lacking challenging scenarios often experienced in everyday driving, from night situations to rain, blinding sun or harsher weather conditions (e.g. snow or fog). Thus, camera-based approaches are favored, while the robustness of a LiDAR is under-represented [19]. In those challenging situations, we expect the UGrid-Fused to be significantly more reliable than the UView-Cam due to the use of multiple redundant sensors based on different measurement methods. Another challenge is that it is not possible to verify a sufficiently correct prediction behavior of the FCN in all possible circumstances. As a consequence, it is helpful to operate multiple methods simultaneously. For instance, an alternative computational method is provided by the LiDAR free-space detection on the OGrid-Fused occupancy grid (cf. Section IV-D), which is also feasible when the camera image is unavailable (e.g. in a sensor malfunction or obstruction). This approach extracts the maximum of the potentially drivable flat ground area, i.e. the drivable space where no collision occurs and could be considered as fallback algorithm.

E. Computational performance

Table I also lists the inference time, i.e. average inference time per input data, of the proposed and related approaches. We achieve an average inference time of $103ms$ and $83ms$ for the *UView-Cam* and *UGrid-Fused* approach respectively. Additionally for the *UGrid-Fused* approach, we have created an optimized C++ implementation for the occupancy grid generation which has a performance of 3.2 million points per second. This results in an average generation time of $35ms$ (assuming a Velodyne HDL-64E with a rotation speed at 10Hz and 180° view area). Thus, both approaches allow for real-time operation at about 10Hz.

VI. Summary and future work

In this work we have proposed two different road detection approaches as foundational building blocks for autonomous vehicles. The UView-Cam approach is based on a single camera and employs a modified VGG19-UNet architecture. In order to obtain a high-quality dataset with binary road labeling, we have unified multiple publicly available datasets and also applied minor manual improvements. Furthermore we have applied an empirical optimization method to identify the best combination of possible parameters, such as pre-trained weights, FCN parameters, colorspace and normalization. To transcend the single sensor view, we also propose the UGrid-Fused approach which fuses camera image and LiDAR scan data into a common multi-dimensional 15 channel occupancy grid representation in BEV. The FCN of the single-camera approach is adapted to work with this occupancy grid as input. The training is done with KITTI data, which both includes camera and synchronized LiDAR data and calibration data.

In an experimental evaluation, we compare the performance of both approaches in the public KITTI urban road detection benchmark, which allows quantitative comparisons to other approaches. For the UView-Cam and UGrid-Fused approach, we have achieved $MaxF$ scores of 94.23% and 93.81% respectively. We also achieved a computational performance of $103ms$ and $83ms$ and therefore a detection rate of about 10Hz. The major drawback of the early fusion is a lack of high quality training data, for which we expect a significant improvement for the UGrid-Fused over the UView-Cam FCN. Hence training and evaluation with data sets of a wider range of harsh weather conditions is an important next step [20].

References

[1] National Highway Traffic Safety Administration et al. Nhtsa (2013). *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices*, 2012.

[2] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[4] J. Fritsch, T. Kuhnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, pages 1693–1700. IEEE, 2013.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Road/lane detection evaluation 2013, 2017.

[6] R. Mohan. Deep deconvolutional networks for scene parsing. *arXiv preprint arXiv:1411.4101*, 2014.

[7] G. Oliveira, W. Burgard, and T. Brox. Efficient deep methods for monocular road segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016)*, 2016.

[8] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *arXiv preprint arXiv:1612.07695*, 2016.

[9] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. *arXiv preprint arXiv:1702.08502*, 2017.

[10] J. Byun, K. Na, B. Seo, and M. Roh. Drivable road detection with 3d point clouds based on the mrf for intelligent vehicle. In *Field and Service Robotics*, pages 49–60. Springer, 2015.

[11] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde. Fast lidar-based road detection using convolutional neural networks. *arXiv preprint arXiv:1703.03613*, 2017.

[12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[14] A. Hermans, G. Floros, and B. Leibe. Dense 3d semantic mapping of indoor scenes from rgb-d images. In *Robotics and Automation (ICRA), 2014 IEEE International Conference*, pages 2631–2638. IEEE, 2014.

[15] J. Alvarez, T. Gevers, Y. LeCun, and A. Lopez. Road scene segmentation from a single image. In *European Conference on Computer Vision*, pages 376–389. Springer, 2012.

[16] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. Lopez. Vision-based offline-online perception paradigm for autonomous driving. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 231–238. IEEE, 2015.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, Enzweiler, et al. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[19] M. Kutila, P. Pyykönen, W. Ritter, O. Sawade, and B. Schäufele. Automotive lidar sensor development scenarios for harsh weather conditions. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 265–270, Nov 2016.

[20] O. Tas, S. Hörmann, B. Schäufele, and F. Kuhnt. Automated vehicle system architecture with performance assessment. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017.