# Semantic Grid-Based Road Model Estimation for Autonomous Driving

Julian Thomas*, Julian Tatsch*, Wim van Ekeren†, Raúl Rojas and Alois Knoll

***Abstract*— For autonomous driving, knowledge about the current environment and especially the driveable lanes is of utmost importance. Currently this information is often extracted from meticulously (hand-)crafted offline high-definition maps, restricting the operation of autonomous vehicles to few well-mapped areas and making it vulnerable to temporary or permanent environment changes. This paper addresses the issues of map-based road models by building the road model solely from online sensor measurements. Based on Dempster-Shafer theory and a novel frame of discernment, sensor measurements, such as lane markings, semantic segmentation of drivable and non-drivable areas and the trajectories of other observed traffic participants are fused into semantic grids. Geometrical lane information is extracted from these grids via an iterative path-planning method. The proposed approach is evaluated on real measurement data from German highways and urban areas.**

## I. Introduction

Developing an accurate, robust and current representation of an autonomous vehicle's environment is one of the major challenges for autonomous driving to become a reality. Higher-level functions like prediction, trajectory- and maneuver planning rely on information about other traffic participants, lane geometry and static obstacles from the road model. Currently many autonomous driving projects rely on meticulously (hand-) crafted offline high definition maps, enriched with sensed traffic participants, traffic light states and static obstacles. This restricts the operation of autonomous vehicles to a few well-mapped areas and makes them vulnerable to temporary or permanent environment changes not yet reflected in the map (e.g. construction sites, freshly applied lane markings). Moreover, the information extracted from the map can only be as good as the accuracy of the localization on the map, which in complex dynamic urban environments still proves to be challenging to guarantee. While the currentness issue could be addressed by a dense fleet of mapping customer vehicles frequently updating the map in the back-end, this inevitably would incur data transmission, data trustworthiness and cost issues.

In contrast to offline high definition maps an online road model can be built solely using sensor measurements directly from the autonomous vehicle. We propose such an online road model built on LiDAR point clouds, lane markings, vehicle trajectories and to the best of our knowledge, for the first time a semantic stereo point cloud.

The remainder of the paper is structured as follows: In section II related work about semantic segmentation, fusing different sensor modalities in grids and building online road models is discussed. In section III an overview of the proposed online road model framework is given. In particular the construction of the grid and the lane extraction are described. In section IV the online road model is evaluated on high definition maps. Section V provides concluding remarks and points out future work.

## II. Related Work

Although the task of online road model estimation from sensor data is not very well studied yet, several lines of work are of relevance for it: camera and LiDAR based lane- and road boundary estimation, detection and tracking of dynamic objects, segmentation of street level images and semantic 3D reconstruction.

Obviously lane markings can and should be used for road model estimation e.g. [1] whenever they are available. However, lane markings are sometimes ambiguous (especially in construction sites), often worn out or the center lines are deliberately removed in urban streets to discourage speeding. In absence of lane markings sometimes at least road boundaries can be estimated from camera or radar measurements (e.g. [2], [3]). Unfortunately, most roads are not clearly delimited by boundaries like guard rails and the curbs are often so low they are indiscernible from automotive-grade laser scans. Also road boundaries only delimit drivable from non-drivable areas without carrying additional information about the drivable area itself. Although road model estimation from road boundaries alone is possible as long as lane parallelism is true (mostly on highways), extracting individual lanes is not as soon as non-parallel lanes in the more complex urban environments must be taken into account. Another valuable source of lane information exists in the trajectories of the other observable traffic participants. Drivable areas and likely lane center lines can be inferred from simply observing the traffic scene. In [4] the trajectories are only used to support the lane boundary estimation of the ego lane. The ego lane itself is estimated based on lane markings detected by a camera system. To sum up, existing road model estimation approaches based on lane marking detection or

* J. Thomas and J. Tatsch are with the Autonomous Driving Project at BMW Group AG, Munich, Germany and contributed equally to the work (julian.(thomas|tatsch)@bmw.de).

†W. van Ekeren is within the department of active ADAS at Altran Deutschland S.A.S. & Co. KG, Munich, Germany (wim.vanekeren@altran.com).

R. Rojas is with the Institut für Informatik, Freie Universität at Berlin, 14195, Berlin, Germany (raul.rojas@fu-berlin.de).

A. Knoll is with the Department of Informatics, Technical University of Munich, 85748, Garching b. München, Germany (knoll@in.tum.de).

road boundary estimation may work well enough for some highway scenarios but are unable to cope with the diversity of rural and urban scenarios. Here the notion of drivable and non-drivable areas is insufficient. Often areas and their semantic meaning cannot be derived from (protruding) geometry alone (e.g. parking spaces, crosswalks, bus- and bicycle lanes). The color, texture and additional cues such as traffic signs are necessary to understand the semantic meaning of these areas and comply with their associated rules. Here the next generation of visual scene understanding is required. Prior scene understanding work like [5], [6] was already concerned with segmenting and 3D reconstructing urban road scenarios from stereo- and mono camera images. However, these approaches were still based on Superpixels and conditional random fields. Then the publication of the CityScapes dataset [7] for the first time offered sufficient training data for a wide variety of classes relevant to autonomous driving, such as road, sidewalk, parking, rail track, vegetation and buildings. This fueled the field of image segmentation methods based on convolutional neural networks (CNNs), e.g. [8]–[14], which outperform the prior work speed- and accuracy-wise. While the segmentation results are often nothing short of astonishing, to be truly useful for autonomous driving, having a spatial context for the semantic information is essential. Spatial information about the environment is typically measured explicitly with ranging sensors such as radar, LiDAR, stereo cameras, inferred explicitly up to scale by structure from motion or inferred implicitly in neural network architectures e.g. [9]. In [15] CNNs are used to classify cells in a radar occupancy grid map into three classes (car, other, unlabeled). [16] propose to first segment an image, feed a top-down view of the segmented image and some LiDAR occupancy maps into a encoder-decoder convolutional neural network which produces a semantic grid image. The most closely related work to ours is [17], which also use a state of the art image segmentation architecture to perform semantic image segmentation and registration to their stereo camera. The semantic and disparity information is fused into an intermediate semantic Stixel representation for improved compactness. In contrast to Stixel-based representations, our work builds on grid-based representations widely used for various tasks in environment perception and situation interpretation. The original idea of storing occupancy probabilities in grid maps [18] has been extended in several directions. In [19] an occupancy grid is built from sensor measurements to extract the road boundaries. In a later publication by the same author the occupancy grid is enhanced by GPS waypoints from an offline map [20]. In a further-reaching approach [21] LiDAR measurements are fused with a-priori knowledge from an offline map in a grid. The work moreover attempts to detect moving objects by analyzing the conflict between the past and the current cell's states (the cell's states change from free to occupied and vice versa). [22] propose a method to generate occupancy grids containing only the static parts of dynamic environments. A particle filter is used to estimate the state of occupancy (occupied by a static or a dynamic obstacle).

In addition to the state of the art, our contribution is twofold. We propose to perform multi-class semantic segmentation directly on the camera image, compute an accurate semantic point cloud and only then fuse the results into a novel semantic evidential grid. We moreover present a probabilistic and easily extensible frame-work based on Dempster-Shafer theory to perform semantic and temporal grid fusion. By using two different beliefs calculated from the fused grid we find dynamically-feasible and collision-free paths and then extract the lane boundaries along these paths yielding the final road model. Fig. 1 shows a schematic overview of the proposed online road model estimation method.

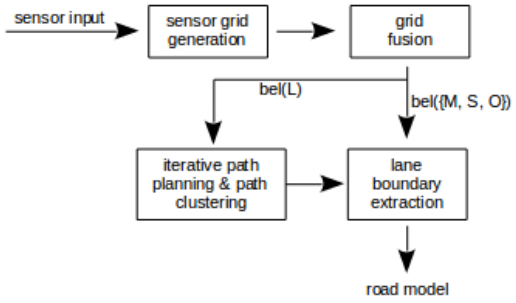



Fig. 1. Schematic overview of the proposed online road model estimation pipeline

## III. Online Road Model Framework

Our road model is defined as a set of lanes

$$\mathcal{R} = \{\mathcal{L}_1, \ldots, \mathcal{L}_n\} \tag{1}$$

where each lane consists of a set of points forming the left and the right boundary respectively.

$$\mathcal{L} = \{\mathcal{B}^{left}, \mathcal{B}^{right}\} \tag{2}$$

with

$$\mathcal{B} = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}, \mathbf{p}_i = (x, y)^T \tag{3}$$

It is important to note that this road model formulation makes no assumptions about the geometrical shapes lanes can have. Additional information such as the type of lane boundaries can easily be added later by extending the definition of $\mathbf{B}$ or $\mathbf{p}$. In contrast to [23] this work is based on the theory of belief functions as brought forward by Dempster and reformulated by Shafer [24]–[26]. Dempster-Shafer theory (DST) is a generalization of Bayesian probability theory and allows to calculate the belief in a specific hypothesis taking all available evidence from different sources into account. In the DST the *frame of discernment* (FOD) is defined as a set $\Omega$, which elements represent all possible states/hypotheses $\theta_i$ of the system under consideration:

$$\Omega = \{\theta_1, \theta_2, ..., \theta_N\} \tag{4}$$

Note, that the elements $\theta_i$ must be mutually exclusive and exhaustive, meaning that at least one hypothesis must be true.

A *basic belief assignment function* (BBA) is used to assign belief masses not only to a single hypothesis $\theta_i$ but to any subset of $\Omega$. Thus, $2^\Omega$ is defined as the set of all subsets of $\Omega$ including the empty set $\varnothing$:

$$2^\Omega := \{U \mid U \subseteq \Omega\} \tag{5}$$

The BBA itself is defined as

$$m : 2^\Omega \rightarrow [0, 1] \tag{6}$$

with the following properties

$$m(\varnothing) = 0 \tag{7}$$

$$\sum_{A \in 2^\Omega} m(A) = 1 \tag{8}$$

The belief mass of an element $A \in 2^\Omega$, written as $m(A)$, is the proportion of all available evidence implying exactly $A$ is true, but no particular subset of $A$. In contrast to that, the belief in $A$, *Bel(A)*, is defined as the sum of the masses of all subsets of $A$ including $A$ itself:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{9}$$

It is the amount of evidence that either the given hypothesis-set $A$ or one of its subsets is true. Belief functions with different BBAs can be combined by combining their respective BBAs. This requires that they are defined over the same frame of discernment. The combination of BBAs directly results in a new belief function, which includes the knowledge/evidences of both belief functions. Let $Bel_1$ and $Bel_2$ two different belief functions over the same frame of discernment, and $m_1$ and $m_2$ their corresponding BBAs, then their BBAs can be combined as follows:

$$m(A) = \frac{1}{1-k} \sum_{A_i \cap B_j = A} m_1(A_i) m_2(B_j), \;\; A \neq \varnothing \tag{10}$$

$$m(\varnothing) = 0 \tag{11}$$

with

$$k = \sum_{A_i \cap B_j = \varnothing} m_1(A_i) m_2(B_j) \tag{12}$$

Eq. 10 is also called *Dempster's Rule of Combination*. $k$ is a measure for the amount of conflict between $m_1$ and $m_2$.

In contrast to other authors, e.g. [27], we propose to define the frame of discernment as

$$\Omega = \{L, M, S, O\} \tag{13}$$

where L stands for *Lane*, M for *Marking*, S for *Sidewalk* and O for *Obstacle*. A lane is defined as the drivable surface between the left and right lane boundary. Markings are the white/yellow lane markings painted on the ground. The hypothesis *Sidewalk* includes the boundary between the lane and the sidewalk as well as the area of the sidewalk itself (not only the curbstone). *Obstacle* comprises any static obstacle which is not traversable. From the power set

$$\begin{aligned} 2^\Omega = \{&L, M, S, O, \{L, M\}, \{L, S\}, \{L, O\}, \{M, S\}, \\ &\{M, O\}, \{S, O\}, \{L, M, S\}, \{L, S, O\}, \{L, M, O\}, \\ &\{M, S, O\}, \varnothing, \Omega\} \end{aligned} \tag{14}$$

we only use a "reduced" power set

$$2_r^\Omega = \{L, M, S, O, \{M, S\}, \{S, O\}, \{M, S, O\}, \varnothing, \Omega\} \tag{15}$$

since in our case the beliefs and masses for all other sets are not measurable with our current sensor setup. Measurement data from different sources is fused and accumulated in a grid-based data representation using the Dempster-Shafer theory and the above mentioned reduced power set. The process is explained in detail in the following section.

### A. Grid-based Information Fusion

A grid is a multidimensional lattice with equally-sized cells, each cell storing stochastic information $m$ inferred from sensor measurements [28]. The belief of a grid $\mathbf{m}$, containing all the probabilistically correct fused sensor measurements, can be written as

$$Bel(\mathbf{m}) = \prod_i Bel(m_i \mid \mathbf{z}_{1:t}, \mathbf{x}_{1:t}) \tag{16}$$

with sensor measurements $\mathbf{z}$ from time 1 up to the current time step $t$, and the ego-vehicle poses $\mathbf{x}$. $m_i$ denotes the cell with index $i$, which is omitted in the following for the sake of better readability. For computational tractability, it is moreover assumed that all cells are conditionally independent of each other, allowing the parallel computation of all cells on the GPU. The ego-motion of the vehicle is compensated by shifting the grids according to the vehicle's pose change $\mathbf{x}_{t_{i+1}} - \mathbf{x}_{t_i}$, which is directly measured by the vehicle's odometry system. Measurements from the lane detection systems, observed object trajectories and semantically segmented point clouds can now be used to infer belief masses for different hypotheses from the reduced power set $2_r^\Omega$. The BBAs are the equivalent to an inverse sensor model $p(m|\mathbf{z}_t)$ given a new sensor measurement $\mathbf{z}_t$ in Bayesian probability theory. The following subsections describe the considerations necessary for fusing the main sensor modalities.

*1) Lane Markings:* Lane markings can either be lane dividers i.e. have the semantic meaning to separate individual lanes, be lane edge lines i.e. delimiting the most outside lane from the emergency lane or road edge lines i.e separating road-regions from non-road-regions. The physical appearance of the lane markings, detected by a state-of-the-art vision system, can at most deliver cues like the type of marking (solid line, dashed line, double solid/dashed line, etc.) as well as its color. The color of lane markings is especially essential for detecting the correct lane markings to follow in construction sites. In our case the lane markings are detected with a trifocal camera system delivering for each marking the type, the color and a set of points $(x, y)$

describing the marking's geometry. The BBA used here to derive the belief masses for different hypotheses from a lane marking measurement is a convolution of a rectangular function with a Gaussian distribution reshaped to values between $[0; 1]$. The evidence of 0 reflects no knowledge about a given cell, whereas 1.0 indicates complete knowledge.

For the belief mass of $M$ (Marking), $m(M)$, the Gaussian models the uncertainty in the position of the measured marking. The rectangular function is used to take the area (on the ground) of the marking itself into account. Therefore, the mean of the Gaussian is positioned in the middle of the detected marking and the width of the rectangular function equals the width of the marking. Depending on the type of the marking the belief mass can also be distributed across other hypothesis groups than $M$ (Marking) only. In our case the lane marking detection system is able to distinguish between marking, sidewalk and road edge. Thus, the mass is either given to $\{M\}$, $\{S\}$ or $\{S, O\}$ in case of a road edge.

Detected markings can also be used to infer belief mass for *Lane*, $m(L)$. For example, for dashed lane markings, because crossing them is allowed, they also indicate the existence of a lane on either side of the marking. Whereas for solid lines, the existence of a lane can only safely be assumed on the vehicle-facing side of the marking. For the lane mass $m(L)$ the mean of the Gaussian is shifted to the left/right by half of a typical lane width in Germany depending on the type of the marking. The width of the rectangular function corresponds to the lane width used for the Gaussian. Since the lane marking sensor measurements are from a single time step $t$ only, we denote the inferred masses as $m_t^{\mathrm{LM}}(\cdot)$ from now on, where the upper index stands for the source of the information - "Lane Marking" in this case.

*2) Moving Objects:* The trajectories of other traffic participants can also be a valuable source for lane and lane boundary information. Traffic participants are detected and tracked in our system using the evidential grid-based tracking algorithm proposed in [29]. The object tracking provides a list of traffic participants with the object's current position taking information from LiDAR, radar and camera sensors into account. Assuming most traffic participants use valid lanes, belief masses for lane and "lane boundaries" can be inferred by using an appropriate BBA. As in the case of lane markings, traffic participants lend itself to be modeled by Gaussian distributions and rectangular functions. For the belief mass of Lane, $m(L)$, the distribution is centered around each measured object position. The width of the rectangular function equals the object's width.

Besides that, it is also possible to infer belief masses for lane boundaries on the left/right side of the object. Since it is not possible to distinguish the type of the boundary, the belief mass is inferred for $\{M, S, O\}$ indicating that it is either *Marking*, *Sidewalk* or *Obstacle*, but there is no further evidence which one exactly it might be. Thus, the belief mass is assigned to $m(\{M, S, O\})$. The mean of the Gaussian is shifted to the left/right of the object's midpoint by half of a typical lane width. As the resulting belief masses

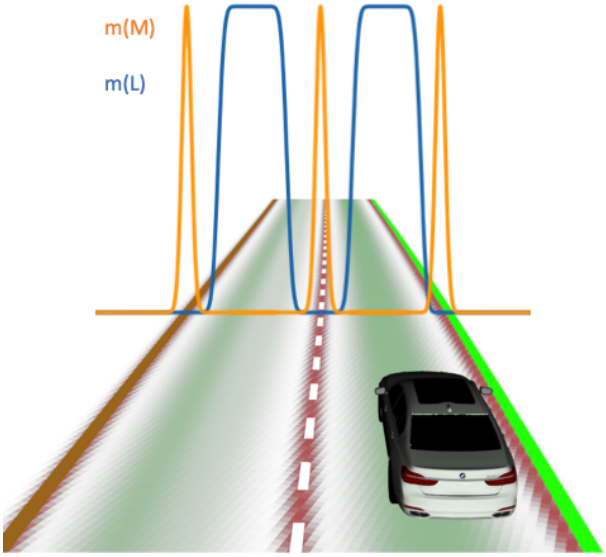



Fig. 2. Situation with 3 markings measured by the lane marking detection system. For the ego lane a dashed marking on the left side and a continuous marking (green line) was detected as well as a continuous marking of the neighbouring lane (brown line). On top, the BBAs for *Marking*, $m(M)$ (orange) and for *Lane*, $m(L)$ (blue), are shown. Below, the resulting grid is displayed, where the color denotes the hypothesis (M: red, Lane: green). The belief mass is encoded in the alpha channel of the grid (transparency).

contain sensor measurements from a single time step - they are denoted by $m_t^{\mathrm{Obj}}(\cdot)$.

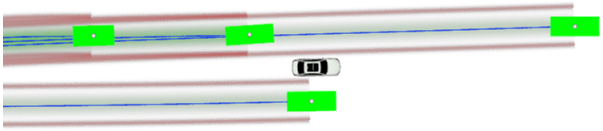


Fig. 3. Grid with belief masses inferred from traffic participant's trajectories. Green boxes are currently measured traffic participants. The blue dots indicate past measurements. The color represents the hypothesis (*Lane*: green, $\{M, S, O\}$: red). The alpha channel (transparency) reflects the mass value.

*3) Semantic Segmentation:* With the recent advent of convolutional neural networks in general and the inception of fully convolutional networks for semantic image segmentation [8] in particular, every pixel in an image can be classified accurately and efficiently by means of an encoder-decoder architecture. The encoder gently reduces the spatial input image resolution with alternating sequences of convolutional and pooling layers, distilling class information - just like in all major image classification architectures. In contrast to these, semantic segmentation networks moreover have a subsequent decoder part with transposed convolution layers. With these they are able to perform learned non-linear upsampling while weaving in spatial cues from skip connections helping to reconstruct the detailed segmentation masks. See Fig. 4 for an urban traffic scene segmented with our re-implementation of PSPNet [11]. While simple monocular

Fig. 4. The input image with a segmentation overlay by our re-implementation of PSPNet. The color of each pixel encodes its class affiliation.

camera images are cheap and abundant, adding a second camera is still affordable and yields a more powerful stereo setup. By determining the disparities between the stereo images e.g. via Semi-Global Matching [30] and computing the depth map, distances to objects in the scene can be measured densely and efficiently. By correctly registering depth and semantically segmented images, a semantic point cloud can be projected into the world coordinate system and is evaluated against our ground truth high definition map. See Fig. 5 for the semantic point cloud overlaid on top of our HD map. To coerce the 3D semantic point cloud into

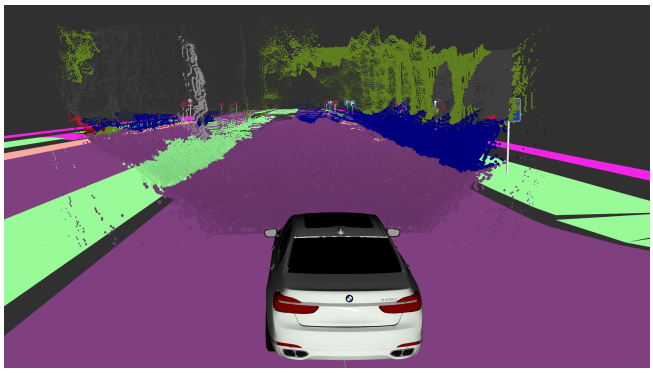


Fig. 5. The semantic point cloud overlaid over our HD map. The color of each pixel encodes its class affiliation.

our common 2D grid representation, all measurements are mapped to their corresponding hypotheses group of the frame of discernment and projected to 2D grids. The derived belief mass for a hypothesis group $A$ at each grid cell is simply the multiplication of the class confidence and the pixel location probability:

$$m(A) = p_{\text{class}}(A) \cdot p_{\text{location}}(A) \tag{17}$$

The class confidence is outputted by the semantic segmentation, whereas the location probability is a result of the stereo disparity calculation. As the semantic segmentation is done for every image frame independently, the resulting grid reflects a single time step only. The resulting grids are therefore denoted as $m_t^{\text{SS}}(\cdot)$.

(a) (b)

Fig. 6. Grid created from the semantic point cloud. (a) Camera image showing the situation. (b) Grid containing the belief masses $m_t^{SS}(\cdot)$. Each pixel represents a grid cell. The color encoding represents the hypothesis (*Lane*: purple, *Sidewalk*: magenta), the alpha channel (transparency) reflects the mass value.

### B. Grid Fusion

In order to obtain belief masses containing all the data the belief masses described in the previous section have to be fused first across all input sources (lane marking, moving objects, semantic segmentation) and later temporally to include all information from the past up to the present time $t$:

$$\mathbf{m}_t^{\text{all}}(\cdot) = \mathbf{m}_t^{\text{LM}}(\cdot) \oplus^{\text{S}} \mathbf{m}_t^{\text{Obj}}(\cdot) \oplus^{\text{S}} \mathbf{m}_t^{\text{SS}}(\cdot) \tag{18}$$

This way, $\mathbf{m}_t^{\text{all}}(\cdot)$ contains the information available from all sensors at a single time step $t$.

For the temporal fusion, the accumulated grid from the previous time step is fused with the current Grid $\mathbf{m}_t^{\text{all}}(\cdot)$:

$$\mathbf{m}_{1:t}^{\text{all}}(\cdot) = \mathbf{m}_{1:t-1}^{\text{all}}(\cdot) \oplus^{\text{T}} \mathbf{m}_t^{\text{all}}(\cdot) \tag{19}$$

Currently, we are experimenting with different implementations of the sensor fusion operator $\oplus^{\text{S}}$ and the temporal fusion operator $\oplus^{\text{T}}$. In this paper, Eq. 10 is used for $\oplus^{\text{S}}$ and the cumulative fusion rule from [31] for $\oplus^{\text{T}}$.

### C. Road Model Extraction

Higher-level functions like prediction and maneuver planning require a continuous and consistent description of the lane geometry and topology graph to work. Different methods to extract these representations from grids were already proposed in [19], [23]. In this paper, lane geometries are extracted by first searching for drivable paths using the belief $Bel(L)$ of $\mathbf{m}_{1:t}^{\text{all}}$ with the path planning method with unknown goal pose described in [32]. It is based on A* [33] and Rapidly-exploring Random Trees (RRTs) [34]. The path extraction is performed as follows:

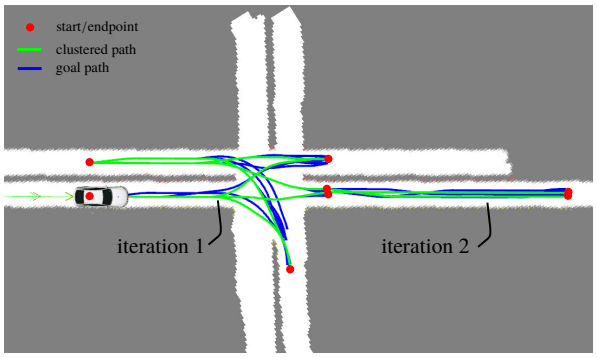



Fig. 7. Example of iterative path planning, showing paths (blue) and clustered paths (green) of two consecutive iterations, with a branching in the first iteration. The underlying lane grid is from a crossing scenario.

1) Sample reasonable starting points around the ego vehicle on the grid
2) For every starting point, find goal paths, using A*/RRT with unknown goal pose
3) Cluster the goal paths
4) Repeat (2) and (3) for $n$ iterations using the end points of the clustered paths as starting point
5) Extract the lane boundaries along the clustered paths from the lane boundary grid

On the grid $\mathbf{m}^{\text{all}}_{1:t}$ containing the belief $Bel(L)$ a dilation with the vehicle's shape as structuring element is performed before using it as cost map. The combination of A* and RRT in the planning with unknown goal pose algorithm ensures that the grid space is explored efficiently. For more details the reader is referred to [35]. The total cost of the path is calculated as a function of the cumulative sum of its cells in the grid along the path. Paths that incur high costs are hence less likely on a lane. The result of the path planner from a single starting point is a set of drivable *paths*. Paths that are believed to be on the same lane are clustered with the method described in [36]. To further improve the efficiency of the planner, it can be run with limited depth and iteratively restarted from the last end points of the clustered paths. Fig. 7 shows iteratively planned paths in a crossing scenario. Due to the random nature of RRT and the noisy costmap, the resulting clustered paths are not smooth enough to be used as lane center lines directly. Instead, for each cell of the clustered path we search orthogonally for left and right lane boundary points using the belief $Bel\{M, S, O\}$ and extract them if they exceed a threshold $\tau_{\text{bnd}}$. Using least squares estimations with smoothness penalties, splines are fitted to the boundary points. Smooth estimations of the lane center line and the lane width are then determined. Additional logic is used to correctly section lanes at branching and merging points and to connect neighboring lanes by association of shared boundaries in the topology graph. Fig. 8 shows an example of the extracted road model using lane markings, semantic segmentation and object trajectories.

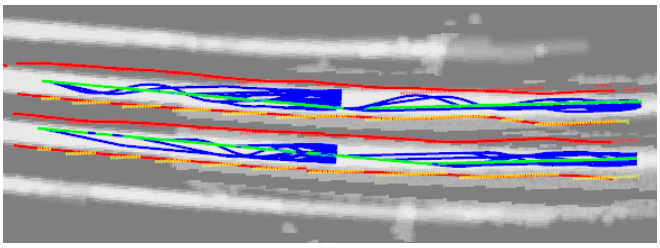


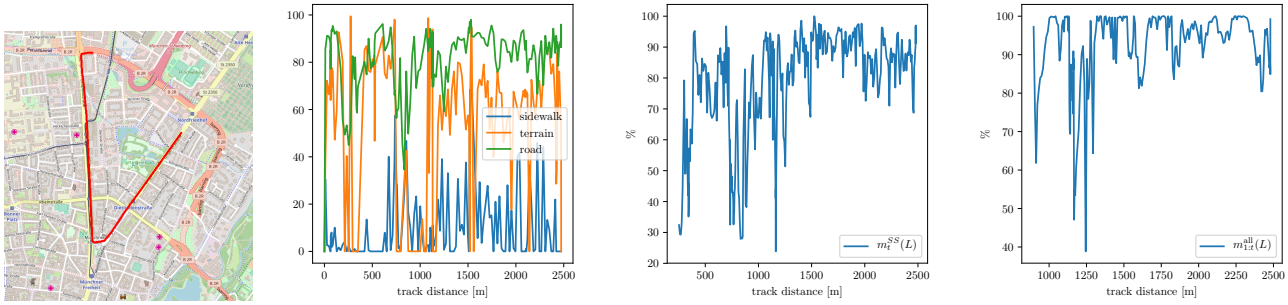
Fig. 8. Two path planners are set up to iteratively plan (starting points on the left side), successfully planned paths are shown in blue, the clustered paths in green, the final lane boundaries are shown as red lines.

## IV. Evaluation

The fused grids and the estimated road model are evaluated on challenging urban scenarios in Munich's inner city and in a commercial area. The test vehicle is equipped with five automotive LiDARs, an automotive lane marking recognition system and a stereo camera. All measurements are processed with our ROS-based autonomous driving framework. To evaluate the presented methods against our globally accurate ground truth HD map, the vehicle localizes itself with high-precision using a Kalman filtered proprietary laser scan matcher and a motion model. Typically, the system achieves an accuracy higher than $\pm$ 0.1 m. Our ground truth HD map was collected exactly the tedious and error-prone way we aim to avoid. The area to be mapped was scanned with high-resolution LiDARs and cameras several times, geometric lane segments were semi-automatically extracted, logically connected and tagged with semantic attributes. To evaluate the accuracy of the semantic point cloud, each semantic point is projected to the HD map and its class is compared with the ground truth. A performance measure is derived by dividing the number of correctly estimated points by the total number of points of a certain class. To evaluate the semantic grids, the appropriate section of the HD map is projected to the grid at the same position and size. The two grids can then be compared on a per cell-basis. For the fused grid, a threshold was used to make a binary comparison per cell. A performance measure is made by dividing the correct grid cells with the total number of grid cells at which the probability exceeds the threshold. Fig. 9 and Fig. 10, show the accuracy of the semantic point cloud and semantic grid for two typical urban scenarios. One can see that especially the road is estimated quite well by the semantic segmentation point cloud and the fused grid.

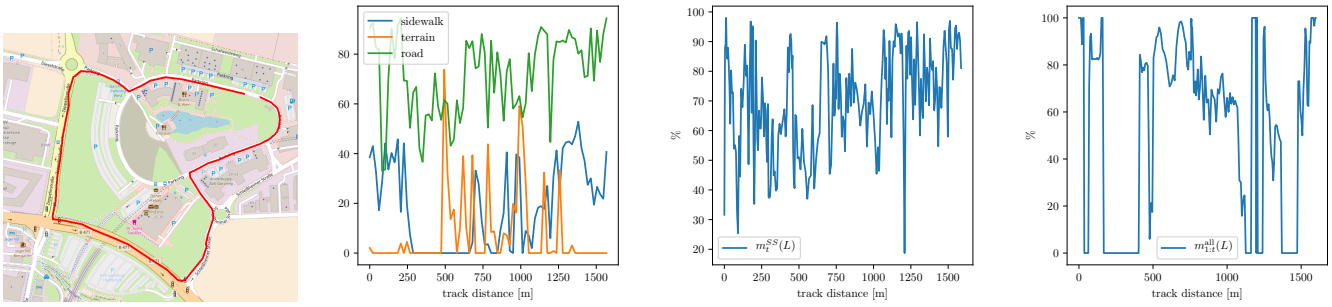
## V. Conclusion

We presented an evidential framework to generate a consistent online road model for autonomous driving. Our approach is solely based on online sensor data and thus completely map and localization independent. Our DST based information fusion framework uses a novel frame of discernment containing all relevant road model elements. Our framework is easily extensible towards new sensor types and can elegantly fuse evidence for subsets of hypotheses in a coherent manner. This allows for the complex multi-class




(a) Track driven (b) Point cloud (c) $m_t^{SS}(L)$ (d) $m_{1:t}^{\text{all}}(L)$

Fig. 9. Evaluated performance over track distance of (a) semantic segmentation point cloud, (b) lane grid from semantic segmentation data and (c) final fused lane grid. The performance indicates the number of correct points or cells as a fraction of the total points or cells.




(a) Track driven (b) Point cloud (c) $m_t^{SS}(L)$ (d) $m_{1:t}^{\text{all}}(L)$

Fig. 10. Evaluated performance over track distance of (a) semantic segmentation point cloud, (b) lane grid from semantic segmentation data and (c) final fused lane grid. The performance indicates the number of correct points or cells as a fraction of the total points or cells.

hypotheses from a semantic point cloud to be fused in a controllable and seamless manner. The resulting semantic grids can also be used for creation or validation of HD maps. The resulting point clouds and grids were also evaluated on real measurement data recorded with one of our autonomous driving prototypes. The evaluation shows that the semantic point cloud is accurate and the resulting fused semantic grid represents the ground truth grid very well. Our implementation of the grid fusion, path planning and road model extraction runs in real-time on GPU hardware.

In future the semantic segmentation part will be optimized in order to let the complete system run in real-time. Enhancing the semantic segmentation by incorporating spatial and temporal cues is currently being investigated. Also, the lane marking polylines accuracy leaves to be desired and with the recent ApolloScape dataset [37], there is now enough data to train for lane marking segmentation and directly integrate it into our segmentation network. Moreover, filtering the resulting road model over time is expected to improve the robustness of the road model estimation considerably. Although first closed-loop tests with our online road model framework have been conducted successfully, it is merely a first step towards mapless road models. Especially semantic relations between extracted lanes and traffic signs and lights remain open field of research.

## References


[1] A. Abramov, C. Bayer, C. Heller, and C. Loy, "A flexible modeling approach for robust multi-lane road estimation," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 2017, pp. 1386–1392.

[2] C. Guo, T. Yamabe, and S. Mita, "Robust road boundary estimation for intelligent vehicles in challenging scenarios based on a semantic graph," in *2012 IEEE Intelligent Vehicles Symposium*, June 2012, pp. 37–44.

[3] C. Adam, R. Schubert, N. Mattern, and G. Wanielik, "Probabilistic road estimation and lane association using radar detections," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, July 2011, pp. 1–8.

[4] C. Guo, J. Meguro, K. Yamaguchi, K. Kidono, and Y. Kojima, "Improved lane detection based on past vehicle trajectories," in *Intelligent Transportation Systems (ITSC), 2014 17th International IEEE Conference on*, 2014, pp. 1956–1963.

[5] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. Torr, "Urban 3d semantic modelling using stereo vision," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 580–585.

[6] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg, "Joint semantic segmentation and 3d reconstruction from monocular video," in *European Conference on Computer Vision*. Springer, 2014, pp. 703–718.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.

[8] 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*, 2015, pp. 3431–3440.

[9] J. Uhrig, M. Cordts, U. Franke, and T. Brox, "Pixel-level encoding and depth layering for instance-level semantic segmentation," in *German*

*Conference on Pattern Recognition (GCPR)*, 2016. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2016/BU16

[10] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *CoRR*, vol. abs/1606.02147, 2016. [Online]. Available: http://arxiv.org/abs/1606.02147

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[12] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," *CoRR*, vol. abs/1704.08545, 2017. [Online]. Available: http://arxiv.org/abs/1704.08545

[13] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Efficient convnet for real-time semantic segmentation," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1789–1794.

[14] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: http://arxiv.org/abs/1802.02611

[15] J. Lombacher, K. Laudt, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic radar grids," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1170–1175.

[16] Ö. Erkent, C. Wolf, C. Laugier, D. Sierra González, and V. R. Cano, "Semantic Grid Estimation with a Hybrid Bayesian and Deep Neural Network Approach," in *IROS 2018 - IEEE/RSJ International Conference on Intelligent Robots and Systems*. Madrid, Spain: IEEE, Oct. 2018, pp. 1–8. [Online]. Available: https://hal.inria.fr/hal-01881377

[17] L. Schneider, M. Cordts, T. Rehfeld, D. Pfeiffer, M. Enzweiler, U. Franke, M. Pollefeys, and S. Roth, "Semantic stixels: Depth is not enough," in *Intelligent Vehicles Symposium (IV), 2016 IEEE*. IEEE, 2016, pp. 110–117.

[18] A. Elfes, "A sonar-based mapping and navigation system," in *Proceedings of the 1986 IEEE International Conference on Robotics and Automation*, vol. 3, 1986, pp. 1151 – 1156.

[19] M. Konrad, M. Szczot, and K. Dietmayer, "Road course estimation in occupancy grids," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*, june 2010, pp. 412 –417.

[20] M. Konrad, M. Szczot, F. Schule, and K. Dietmayer, "Generic grid mapping for road course estimation," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*, June 2011, pp. 851–856.

[21] M. Kurdej, J. Moras, V. Cherfaoui, and P. Bonnifait, "Map-aided evidential grids for driving scene understanding," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 1, pp. 30–41, Spring 2015.

[22] G. Tanzmeister, J. Thomas, D. Wollherr, and M. Buss, "Grid-based mapping and tracking in dynamic environments using a uniform evidential environment representation," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 6090–6095.

[23] J. Thomas, K. Stiens, S. Rauch, and R. Rojas, "Grid-based online road model estimation for advanced driver assistance systems," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, June 2015, pp. 71–76.

[24] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[25] ——, "A generalization of bayesian inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 205–247, 1968.

[26] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[27] M. Kurdej, J. Moras, V. Cherfaoui, and P. Bonnifait, "Map-aided fusion using evidential grids for mobile perception in urban environment," in *Proceedings of the 2nd International Conference on Belief Functions*, ser. Advances in Intelligent and Soft Computing, M.-H. Denoeux, Thierry; Masson, Ed., vol. 164. Compiègne, France: Springer, May 2012, pp. 343–350. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00714465

[28] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46 –57, 1989.

[29] S. Steyer, G. Tanzmeister, and D. Wollherr, "Object tracking based on evidential dynamic occupancy grids in urban environments," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1064–1070.

[30] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 807–814.

[31] A. Jøsang, J. Diaz, and M. Rifqi, "Cumulative and averaging fusion of beliefs," *Information Fusion*, vol. 11, no. 2, pp. 192–200, 2010.

[32] G. Tanzmeister, M. Friedl, D. Wollherr, and M. Buss, "Path planning on grid maps with unknown goal poses," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, Oct 2013, pp. 430–435.

[33] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[34] S. M. LaValle, "Rapidly-exploring random trees: A new tool for path planning," 1998.

[35] G. Tanzmeister, M. Friedl, D. Wollherr, and M. Buss, "Efficient evaluation of collisions and costs on grid maps for autonomous vehicle motion planning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2249–2260, Oct 2014.

[36] G. Tanzmeister, D. Wollherr, and M. Buss, "Environment-based trajectory clustering to extract principal directions for autonomous vehicles," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2014, pp. 667–673.

[37] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," *arXiv preprint arXiv:1803.06184*, 2018.