2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Automatic Mapping of Tailored Landmark Representations for Automated Driving and Map Learning

Jan-Hendrik Pauls[†,*,1], Benjamin Schmidt[*,1], and Christoph Stiller[1].

***Abstract*— While the automatic creation of maps for localization is a widely tackled problem, the automatic inference of higher layers of HD maps is not. Additionally, approaches that learn from maps require richer and more precise landmarks than currently available.**

In this work, we fuse semantic detections from a monocular camera with depth and orientation estimation from lidar to automatically detect, track and map parametric, semantic map elements. We propose the use of tailored representations that are minimal in the number of parameters, making the map compact and the estimation robust and precise enough to enable map inference even from single frame detections. As examples, we map traffic signs, traffic lights and poles using upright rectangles and cylinders.

After robust multi-view optimization, traffic lights and signs have a mean absolute position error of below 10 cm, extent estimates are below 5 cm and orientation MAE is below 6°. This proves the suitability as automatically generated, pixel-accurate ground truth, reducing the task of ground truth generation from tedious 3D annotation to a post-processing of misdetections.

I. Introduction

For the past decades, maps in robotics were mostly associated with localization, leading to the famous problem of simultaneous localization and mapping (SLAM) [1]. However, many other applications in mobile robotics, like scene understanding and decision making, benefited equally from knowing the ego pose in a rich semantic map, nowadays often called high definition (HD) map. While typical SLAM approaches tackle the automatic creation of localization-focused maps, the automatic creation of a semantic HD map for other applications is not yet solved.

A. HD Maps for Automated Driving

In the context of automated driving, HD maps contain routing graphs, road geometries, but also elements that introduce traffic rules or dictate specific behavior, such as road signs or traffic lights [2].

The automatic mapping of road geometries and markings has often been tackled [3], [4]. Similar mapping approaches for other map elements, such as traffic lights or signs, either lack a rich representation or build upon expensive mobile mapping equipment. The inference of traffic rules and behavior decisions, however, is only possible by accurately knowing the positions and orientations of traffic signs and traffic lights [5]. This raises the need for approaches that are able to estimate those map elements accurately using only on-board sensors of automated vehicles.

Fig. 1. Top: Examplary mapped representations are reprojected into the image with colors depicting the semantic classes. Bottom: 3D view of the same scene.

At the same time, a fleet of cars with highly accurate mapping capabilities could solve street maintenance or urban inventory management tasks that currently require extra work or equipment.

B. Learning from HD Maps

From a machine learning perspective, detailed semantic HD maps can be used to easily generate the massive amounts of data that will be necessary to safely perceive the static world. Together with pixel-accurate localization and occlusion detection, map elements can be reprojected into sensor data and serve as ground truth. This method, hereafter called *map learning*, however, requires landmarks representations that are rich enough to be rendered correctly as bounding box, possibly even including mask information.

C. Contribution and Outline

We show that the combination of lidar point measurements and semantically segmented camera images is sufficient to extract, track and estimate a highly accurate parametric representation of semantic landmarks for both automated driving and map learning (cf. Fig. 1). More precisely, our contributions are the following:

[†]Corresponding author, jan-hendrik.pauls@kit.edu
[*]The authors contributed equally.
[1]Institute of Measurement and Control Systems, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany.

- We propose to use minimal parametric, *tailored* representations that match the properties that can be expected given the semantic class of certain map elements.
- We show how to fuse semantic instance detections from pretrained neural networks and lidar data to enable highly precise depth and orientation estimation from a single measurement frame (Section III).
- Section IV describes a data association method that builds upon the precise single frame measurements, allowing to combine at least 180 moving frames.
- We propose a robust multi-view landmark estimation (Section V) and demonstrate in the evaluation (Section VI) that our approach can achieve the spatial, extent and orientation accuracy similar to human annotations.

While the basic idea is not new, we show how to combine modern methods like instance segmentation with the idea of a minimal parametric representation to a robust mapping framework.

Recall and precision are mainly dominated by the detection performance of the neural network. Our approach, however, leverages tailored representations and spatial consistency to reduce the problem of manual map or ground truth creation to collecting data and deleting false positive detections during post-processing. Expensive and time-consuming 3D labeling work is thus reduced to deleting false detections, a task similar in effort to semi-automatic 2D object annotation.

II. Related Work

To create HD maps for automated driving, many approaches have tackled the mapping of road surface [3], [4], [6], but also poles [7]–[9] or a combination thereof and planes [10]. While some used a cylindrical model for detection, only [8] stored it and just used it for data association. We incorporate poles as map elements since they are often used as landmarks for localization, but use the diameter to correctly render a bounding box. Also, using semantics, we are able to distinguish poles from trees. Still, poles as landmarks are actually not necessary for automated driving, unlike traffic lights and signs which we would like to put our focus on.

Other work explicitly covered traffic sign mapping. However, instead of using representations that are interpretable or usable for map learning, they focus on a mere position, but often have a more detailed class. Approaches use only cameras [8], [11], mobile mapping lidars [12], deep neural networks [13], [14] or a combination thereof [15], [16]. The last two are closest to our approach, but either have a very complex representation or simply model the position of a sign. Instead, we propose to use a tailored parametrization that lies in between and can still be estimated precisely using lidar sensors made for automated driving instead of mobile mapping. In [17], multiple views are used to estimate shape and pose, achieving high accuracy using very specific class-dependent shape knowledge. As resolving the semantic class of a sign is a problem of its own, we instead propose the use of generic lidar measurements. [18] pursue an idea similar to our approach for sign detection only: The semantic class of a traffic sign is combined with a tailored, minimal 2D shape representation, both detected from camera images using a neural network. For further work about sign detection (without mapping), we refer to [19], [20].

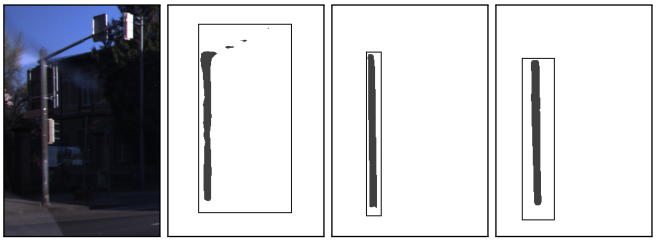

Fig. 2. Three duplicate detections of the same pole shown with their bounding box and mask. Only the second detection is processed further.

On the area of traffic light mapping, most approaches [21]–[25] use cameras as sensor modality, often focusing more on the traffic light status than on accurate pose or position. Estimating the orientation as well as the status would be interesting extensions to our work. For further work on detecting traffic lights, we refer to [26]. To the best of our knowledge, for traffic lights, the combination of lidar and deep neural networks applied on camera images has not been tried, yet. For both, signs and traffic lights, we furthermore propose to use instance masks instead of bounding boxes, improving the mapping of non-rectangular signs.

Finally, another possibly closely related topic is object-based SLAM. Previous approaches focused on generic representations, such as centroids [27], quadrics [28], [29], cuboids [30] or dense objects [31]. In contrast, we propose to make representations as compact and as robust as possible by applying expert knowledge about a minimal parametric representation.

III. Parametric Detections

As preliminary steps of our approach, we apply a pre-trained panoptic neural network [32] to obtain instance masks and bounding box detections of various semantic classes, including road signs, traffic lights and poles. Also, we apply a highly accurate visual SLAM approach [33] to obtain pose estimates and focus on mapping for now.

Given semantic detections $\mathcal{D}_k$, ego motion-compensated lidar points $\mathcal{L}_k$ and pose estimates for each frame k, we obtain parametric measurements via a pre-filtering and a robust depth estimation step.

A. Pre-filtering

Despite non-maxima suppression, the panoptic network occasionally detects the same landmark multiple times, leading to duplicate landmarks that need to be filtered out. Hence, we calculate the mask intersection over union (IoU) between every detection in $\mathcal{D}_k$. Two masks are considered duplicates if their IoU is greater than 10 %. As can be seen in Figure 2, usually the mask with the highest fill ratio is desired and thus kept as true detection.

Our measurement method assumes that the bounding box roughly describes its object's contours. Since there are many

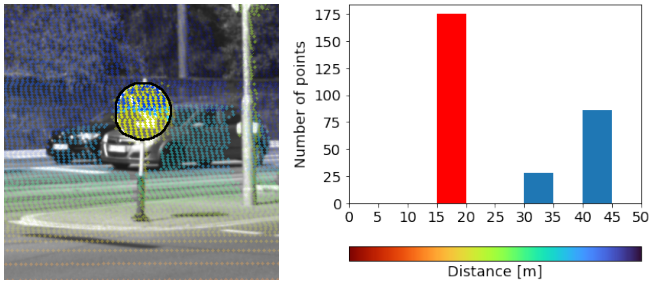

Fig. 3. On the left, lidar points are projected into the cropped image (originally RGB) with depth encoded according to the color map on the right and the mask contour in black. Only the yellow points actually belong to the traffic sign, showing an artifact due to parallax.
On the right, we show the corresponding depth histogram of all points with the cluster selected by our approach marked in red.

poles with arms not fitting our parametric model, we also require a mask fill ratio of at least 30 %. For the same pole, this prefers detections without arms over of those with arms (cf. Fig. 2).

B. Depth Estimation

For highly accurate depth and orientation estimation, we rely on lidar points. Hence, we find the relevant lidar points of the m-th detection $\mathcal{L}_k^m \subseteq \mathcal{L}_k$. We project $\mathcal{L}_k$ into image space using a spherical camera model [34]. The model's as well as extrinsic calibration parameters are retrieved through the methods of [35]–[37]. We then obtain $\mathcal{L}_k^m$ by filtering the projected $\mathcal{L}_k$ with the detection's instance mask.

Distinct mounting positions and motion during recording induce a noticeable parallax effect whereby distant lidar points move onto closer objects in the camera's perspective (cf. Fig. 3). Thus, we cluster the lidar points in the bird's eye view using the DBSCAN algorithm [38] with 2D Euclidean distance, $\epsilon = 0.4\,\mathrm{m}$ and at least 2 points per cluster. We only process the cluster with the closest centroid.

C. Parametric Measurements

Finally, we fuse semantic detections $d_m \in \mathcal{D}_k$ with corresponding lidar points $\mathcal{L}_k^m \subseteq \mathbb{R}^3$ to obtain tailored parametric measurements. Semantic detections consist of a class label c_m and top-left/bottom-right bounding box corners in image coordinates $\mathbf{d}_{\text{TL/BR}}^m \in \mathbb{R}^2$.

The idea of our parametric measurements is that they are tailored to suit the semantic classes (cf. Fig. 4). For poles and traffic lights, we assume an upright standing cylinder which is parameterized by its center point $\mathbf{x} = (x, y, z)^T \in \mathbb{R}^3$, its width or diameter w, and its height h. For traffic signs, we assume an upright standing rectangle which has an additional orientation angle φ around the *up* or z axis.

First, a centroid $\mathbf{x}_\mathcal{L}$ of $\mathcal{L}_k^m$ is calculated robustly by solving the following non-linear least squares system

$$\mathbf{x}_{\mathcal{L},\theta} = \underset{\mathbf{x}_{\mathcal{L},\theta}}{\operatorname{argmin}} \sum_{\mathbf{l}_i \in \mathcal{L}_k^m} \rho\left(\|\mathbf{l}_{i,\theta} - \mathbf{x}_{\mathcal{L},\theta}\|^2\right) \tag{1}$$

for $\theta \in \{x, y, z\}$ with Cauchy loss at scale $a = 0.25$

$$\rho(s) = a^2 \log\left(1 + \frac{s}{a^2}\right). \tag{2}$$

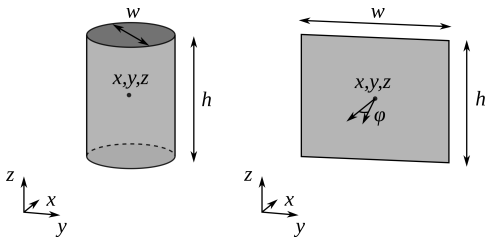

Fig. 4. Two exemplary tailored parametric models: a cylinder for traffic lights and poles as well as an upright rectangle oriented around the *up* axis for traffic signs.

We found that the actual center point $\mathbf{x}$ usually lies close to the bounding box's center viewing ray $\mathbf{d}_C$ while the lidar points, especially with a low number of lidar rays, are generally not evenly distributed around it. To calculate the actual center point $\mathbf{x}$, we intersect an upright plane with $\mathbf{d}_C$ as depicted in Figure 5.

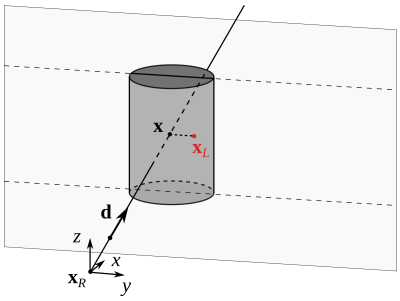

Fig. 5. The robust lidar centroid $\mathbf{x}_L$ is projected onto the center viewing ray $\mathbf{d}_C$ to obtain a 3D center estimate.

The upright plane is given through the lidar centroid $\mathbf{x}_\mathcal{L}$ for the depth and the projection of $\mathbf{d}_C$ onto the ground, $\mathbf{d}_G$, as normal vector.

$$\mathbf{n} = \mathbf{d}_G = \frac{1}{\sqrt{d_{C,1}^2 + d_{C,2}^2}} \begin{pmatrix} d_{C,1} \\ d_{C,2} \\ 0 \end{pmatrix} \tag{3}$$

Thus, we can compute the center point as

$$\mathbf{x} = \frac{\mathbf{x}_\mathcal{L}{}^T \mathbf{n}}{\mathbf{d}_C{}^T \mathbf{n}} \mathbf{d}_C = \frac{\mathbf{x}_\mathcal{L}{}^T \mathbf{d}_G}{\mathbf{d}_C{}^T \mathbf{d}_G} \mathbf{d}_C \tag{4}$$

To obtain w and h, we relate the bounding box's corner points to 3D space. For cylindrical representations, we calculate the 3D corner points by intersecting the viewing rays to the respective corner points $\mathbf{d}_{\text{TL/BR}}$ with the same plane as for determining $\mathbf{x}$:

$$\mathbf{x}_{\text{TL/BR}} = \frac{\mathbf{x}^T \mathbf{d}_G}{\mathbf{d}_{\text{TL/BR}}^T \mathbf{d}_G} \mathbf{d}_{\text{TL/BR}} \tag{5}$$

For traffic signs, we calculate the 3D corner points with the sign's estimated normal vector $\mathbf{n}_\mathcal{L}$, obtained by a total least squares line fit through the lidar points:

$$\mathbf{x}_{\text{TL/BR}} = \frac{\mathbf{x}^T \mathbf{n}_\mathcal{L}}{\mathbf{d}_{\text{TL/BR}}^T \mathbf{n}_\mathcal{L}} \mathbf{d}_{\text{TL/BR}} \tag{6}$$

The normalized estimated normal vector $\mathbf{n}_\mathcal{L}$, like $\mathbf{d}_G$, is projected onto the ground and allows to calcuate the orientation

Class	α_{xs}	α_x	α_w	α_h
Traffic signs	4	3.5	2	2
Traffic lights	4	3.5	2	2
Poles	8	8	4	1

TABLE I. Association cost weights, chosen empirically.

φ through the normalized negative driving direction $-\mathbf{d}_D$:

$$\varphi = \arccos\left(-d_D^T \mathbf{n}_{\mathcal{L}}\right) \tag{7}$$

Finally, the width w and height h are calculated from the 3D corner points' x/y and z coordinates, respectively.:

$$w = \sqrt{(x_{TL} - x_{BR})^2 + (y_{TL} - y_{BR})^2} \tag{8}$$

$$h = z_{TL} - z_{BR} \tag{9}$$

D. Bounding Box Measurements

Lidar resolution decreases quadratically with distance. Thus, for detections with $|\mathcal{L}_k^m| < 5$, we use image-only bounding box measurements which we only use to refine the center point $\mathbf{x}$, but neither update any other parameters nor create new landmarks. Like before, we intersect the $\mathbf{d}_C$ with the already estimated plane, yielding a 3D point. This especially helps for lidar sensors with fewer rays.

IV. Data Association

While our approach works in both directions, iterating over all frames against the driving direction, i.e. backwards in time, helps especially with less powerful lidars to initialize an estimate in close range with high sensor resolution in both lidar and camera. As landmarks are moving away, additional measurements are then associated to already mapped and optimized parametric landmarks using the Hungarian algorithm [39], [40]. Known classes allow to only associate semantically identical objects. Due to extra parameters such as height and width, data association becomes reliable even in cluttered environments with many map elements.

The Hungarian algorithm solves a matrix with cost values between each measurement and landmark. The measurement's parameters are transformed to their respective landmark's coordinate frame using the given poses. The cost to associate landmark ℓ_i with parametric measurement p_j with equal classes is calculated as follows:

$$J_{i,j} = \alpha_{xs}\|\Delta\mathbf{x}\|^2 + \alpha_x\|\Delta\mathbf{x}\| + \alpha_w\Delta w + \alpha_h\Delta h \tag{10}$$

We set the weighting values according to Table I.

To cope with false or new detections, we added backup entries with gating costs of 50 to the cost matrix. Therefore, no measurement is associated if its association cost is higher. A new landmark is created for all parametric measurements without association by taking the measurement as initial estimate $\ell_i = p_j$.

V. Map Optimization

The previously described measurements and data association allow to observe more than 180 moving frames for a single landmark. Hence, in the map optimization step, optimal landmark estimates $\hat{\ell}_i$ need to be calculated from the set of all associated parametric measurements $p_j \in \mathcal{A}_{\hat{\ell}_i}$. To improve data association, we optimize the map whenever a landmark got associated with a new measurement.

For map optimization, we estimate each parameter $\theta \in \{x, y, z, w, h, \varphi\}$ by solving a non-linear least squares with the same robust loss function as previously (cf. Equation (2)):

$$\hat{\ell}_{i,\theta} = \underset{\ell_{i,\theta}}{\operatorname{argmin}} \sum_{p_j \in \mathcal{A}_{\hat{\ell}_i}} \rho\Big(\|\ell_{i,\theta} - p_{j,\theta}\|^2\Big) \tag{11}$$

Of course, the difference in orientation angle φ is only taken into account for traffic signs. Angular wrapping problems never posed a problem, but might do so when aggregating multiple drives.

VI. Evaluation

Unfortunately, there is no suitable mapping benchmark and our neural network turned out to not generalize well to older cameras like those used in the KITTI benchmark [41].

Hence, for evaluation, we map three challenging parts of a larger test lap in the German city of Karlsruhe with our measurement vehicle. As sensors, we employ a global shutter color camera with 4096×1536 pixels that is triggered with 10 Hz when the Velodyne VLS-128 Alpha Prime lidar passes the center of the image. To avoid redundancy, we omit all frames in which the car did not move.

Single Measurement Precision

In Figure 6, a box plot with whiskers at 1.5 IQR shows the precision of the tailored parametric measurements. Coordinates x, y, z are relative to the vehicle pose when the landmark was detected for the first time.

While the positional precision is similar to previous approaches [10], note the precision of width and height measurements as well as the very precise orientation measurements for traffic signs. The position and orientation accuracy implies that for the needs of automated driving, like localization or map inference, already single frame measurements are usually sufficient.

Map Optimization Results

To evaluate not only precision, but also accuracy, we manually labeled the three sections by observing both reprojections in the camera images and aggregated lidar point clouds. As basis we render the automatic mapping outputs for each section using RViz interactive markers [42]. We matched the surface of each marker to its corresponding lidar points and adjusted width/diameter and height using its reprojection in the camera images. For every missing landmark, we duplicated an existing landmark and repeated the previous adjustment steps. Furthermore, this allows to evaluate detection metrics, such as recall and precision.

Spatial accuracy is listed in Table II and depicted in Figure 7 like for the single measurements. For traffic lights and road signs, position estimates are below 10 cm mean absolute error (MAE) while poles suffer from their z coordinate being not well-observable. This problem also impairs the height estimate for poles and can only be corrected when

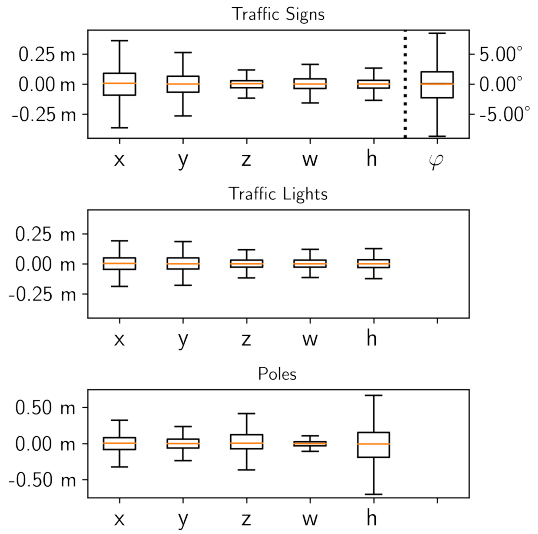

Fig. 6. Precision of single frame measurements. For most automated driving applications, like traffic rule inference, no tracking or aggregation is needed.

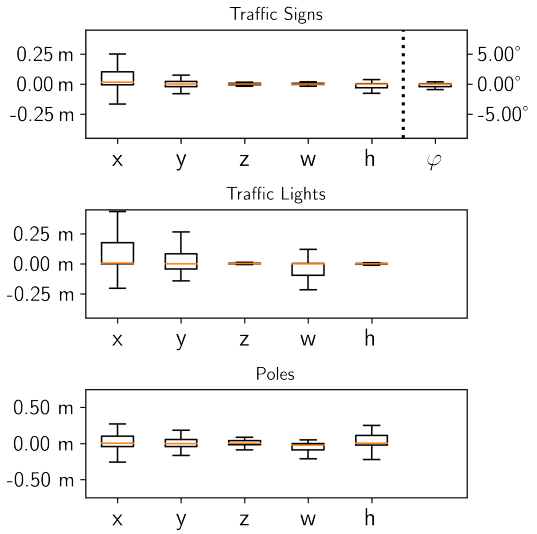

Fig. 7. Accuracy of the optimized map elements, measured against manually labeled ground truth. Position estimates suffer from pose drift, but extent and orientation accuracy is close to human performance.

observability of the pole at both the upper and lower ends is explicitly modeled.

Position accuracy is limited by a slight drift in the poses used for alignment, especially during curves. We would have expected a small depth bias in x-direction for poles or traffic lights since we estimate depth on the outside instead of their center. However, this is not noticeable comparing their depth accuracy with that of unbiased traffic signs.

For signs and traffic lights, extent estimates are in the same order of magnitude that we found achievable during human annotation. The same holds for orientation estimation which, even using an aggregated point cloud, is close to if not even as accurate as human performance.

Data retrieval metrics are listed in Table III. We evaluated two measures: First, we evaluated all traffic lights and signs that are relevant for the road we drove on, called *ego-road*. This measures the applicability for automated driving. Second, we measured recall and precision for all visible elements and evaluated both recall and precision to measure the need for human correction if used to generate ground truth for map learning.

Over all three sequences, considering the ego-road, we were able to map all but one road sign and all but one traffic light. Both were only visible for a few frames.

For map learning, we suffer from consistent false detections like when advertisement signs are classified as road signs. Also due to misdetections from the pretrained neural network, we are missing some poles and suffer in terms of precision as trees and corners of buildings are identified as poles. Thus, while our method has the advantage of being able to distinguish between poles and trees, it would profit from being combined with lidar-based pole detection methods [9].

However, given the spatial, extent and orientation accuracy as well as recall along the ego road, our approach mainly reduces the task of 3D annotation to a removal of false positive detections. Missing detections could even be added manually like a semantic detection from a neural network.

Qualitative results can be found in Figure 8. The robustness of the data association is visualized in Figure 9.

Ablation Study

Finally, we provide an ablation study. Since the neural network showed to be fairly robust against scaling down and deteriorating the images, we only evaluate the consequences of less capable lidar sensors. To simulate a lidar sensor with 64, 32 and 16 layers, we only used every second, fourth or eighth layer. To get closer to the actual performance of most 16 ray lidars, we artificially limit the range to 80 m in case of using 16 layers.

The results in Tables IV and V show that annotation quality is largely independent of the number of beams. Recall, however, suffers with decreasing number of beams and we were not able to compensate for it using bounding box measurements. For automated driving tasks, this holds especially for the 32 beam and range-limited 16 beam variants.

	x	y	z	w	h	φ
Traffic Signs	0.09	0.07	0.03	0.03	0.06	5.6
Traffic Lights	0.11	0.08	0.03	0.04	0.03	-
Poles	0.12	0.10	0.33	0.06	0.62	-

TABLE II. Spatial, extent and orientation accuracy measured as mean absolute error (MAE) in meters or degrees, respectively.

	Recall ego-road	Recall all	Precision all
Traffic signs	97 %	77 %	77 %
Traffic lights	92 %	80 %	83 %
Poles	-	75 %	52 %

TABLE III. Retrieval metrics evaluated for the case of automated driving (ego-road) and automated ground truth generation (all visible).

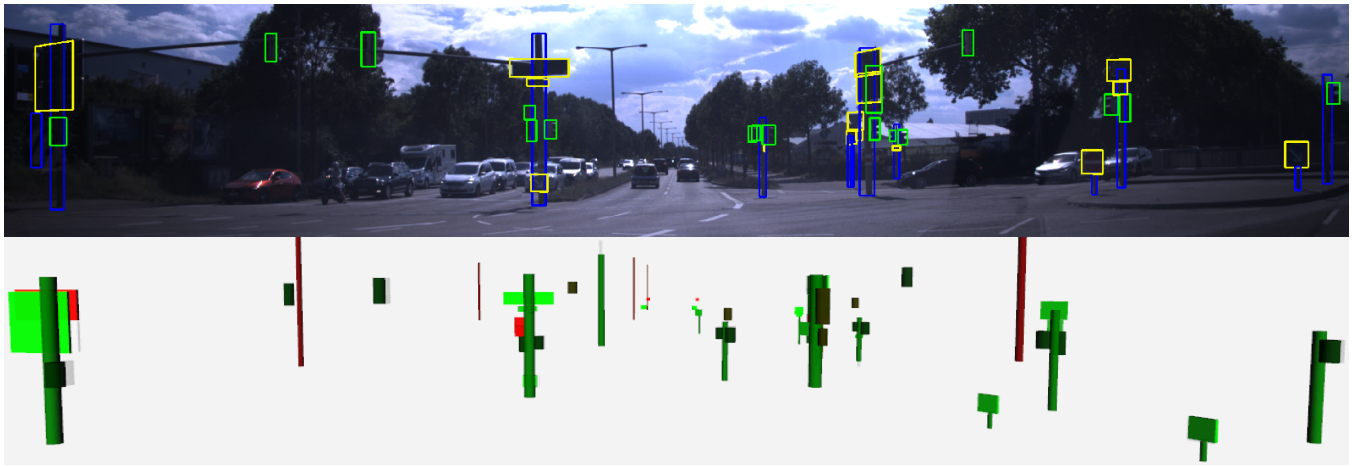

Fig. 8. Qualitative results. Top: The mapped elements are rendered into an image with colors depicting semantic classes. Bottom: 3D view of the scene. True positives are green, false negatives are red and ground truth is gray.

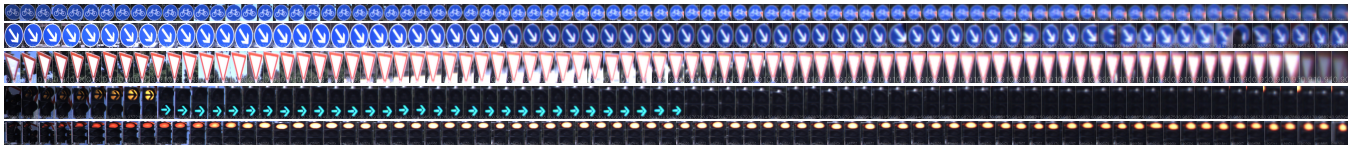

Fig. 9. Qualitative visualization of the data association. Three traffic signs and two traffic lights with their corresponding image crops.

	# Rays	x	y	z	w	h	φ
Traffic Signs	128	0.09	0.07	0.03	0.03	0.06	5.6
	64	0.09	0.07	0.04	0.04	0.05	6.0
	32	0.11	0.08	0.04	0.03	0.06	3.4
	16*	0.12	0.09	0.09	0.05	0.09	4.5
Traffic Lights	128	0.11	0.08	0.03	0.04	0.03	-
	64	0.11	0.08	0.03	0.04	0.03	-
	32	0.11	0.08	0.03	0.05	0.04	-
	16*	0.09	0.08	0.07	0.07	0.05	-
Poles	128	0.12	0.10	0.33	0.06	0.62	-
	64	0.14	0.10	0.40	0.07	0.65	-
	32	0.14	0.11	0.38	0.08	0.65	-
	16*	0.14	0.11	0.36	0.09	0.67	-

TABLE IV. Ablation study on spatial, extent and orientation accuracy over varying beam number showing mean absolute error (MAE). Units are in meters or degree, respectively. *Range limited to 80 m (see text).

	# Rays	Recall ego-road	Recall all	Precision all
Traffic Signs	128	97 %	77 %	77 %
	64	97 %	73 %	83 %
	32	91 %	61 %	88 %
	16*	86 %	50 %	95 %
Traffic Lights	128	92 %	80 %	83 %
	64	92 %	73 %	82 %
	32	85 %	62 %	90 %
	16*	70 %	40 %	97 %
Poles	128	-	75 %	53 %
	64	-	70 %	61 %
	32	-	59 %	76 %
	16*	-	49 %	85 %

TABLE V. Ablation study on recall and precision for different lidar variants. *Range limited to 80 m (see text).

VII. Conclusion

We proposed to use tailored representations for map elements, such as poles, traffic lights or road signs. They enable orientation estimation for signs which is important for automated map generation. Furthermore, for map learning, they are the minimal representation to render bounding boxes and, thus, can serve to automatically generate ground truth from maps. Being minimal in the number of parameters also improves accuracy and robustness. For many other semantic classes, such as manholes or road markings, similar tailored representations can be found.

We showed how to directly measure such tailored representations using lidar points and a pretrained instance segmentation network applied on camera images, allowing data association over more than 180 moving frames. Moreover, they enable the use for automated driving tasks, such as map inference, from even a single frame.

After robust optimization, our estimates are close to human labeling performance. Hence, the proposed approach reduces the task of ground truth generation from tedious 3D annotation to merely filtering out false positive detections.

To improve the results even further, in future work, we will optimize both, the original poses and the landmarks in a joint pose graph optimization. This will enable localization and landmark estimation across different drives. Additionally, we will robustly aggregate the instance mask information.

With pixel-accurate localization building upon previous work [43], the presented approach allows to automatically generate a map that is compact in storage, but still rich enough to serve as instance segmentation or object detection ground truth with vastly reduced mapping effort. Thus, this is an important step towards self-improving SLAM.

References

[1] S. Thrun, W. Burgard, D. Fox, and R. C. Arkin, *Probabilistic Robotics*. MIT Press, 2005.

[2] F. Poggenhans, J.-H. Pauls, J. Janosovits, S. Orf, M. Naumann, F. Kuhnt, *et al.*, "Lanelet2: A high-definition map framework for the future of automated driving", in *2018 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, 2018, pp. 1672–1679.

[3] M. Schreiber, C. Knöppel, and U. Franke, "LaneLoc: Lane marking based localization using highly accurate maps", in *2013 IEEE Intell. Vehicles Symposium (IV)*, 2013, pp. 449–454.

[4] F. Poggenhans, M. Schreiber, and C. Stiller, "A Universal Approach to Detect and Classify Road Surface Markings", in *2015 IEEE 18th Int. Conf. Intell. Transp. Syst.*, 2015, pp. 1915–1921.

[5] F. Poggenhans, "Generierung hochdetaillierter Karten für das automatisierte Fahren", PhD thesis, Faculty of Mechanical Engineering, Karlsruhe Institute of Technology, 2019.

[6] J. Jeong, Y. Cho, and A. Kim, "Road-SLAM : Road marking based SLAM with lane-level accuracy", in *2017 IEEE Intell. Vehicles Symposium (IV)*, 2017, pp. 1736–1473.

[7] R. Spangenberg, D. Goehring, and R. Rojas, "Pole-based localization for autonomous vehicles in urban scenarios", in *2016 IEEE/RSJ Int. Conf. Intell. Robots and Systems (IROS)*, 2016, pp. 2161–2166.

[8] M. Sefati, M. Daum, B. Sondermann, K. D. Kreisköther, and A. Kampker, "Improving vehicle localization using semantic and pole-like landmarks", in *2017 IEEE Intell. Vehicles Symposium (IV)*, 2017, pp. 13–19.

[9] A. Schaefer, D. Buscher, J. Vertens, L. Luft, and W. Burgard, "Long-Term Urban Vehicle Localization Using Pole Landmarks Extracted from 3-d Lidar Scans", in *2019 European Conference on Mobile Robots (ECMR)*, Prague, Czech Republic, 2019, pp. 1–7.

[10] J. Kümmerle, M. Sons, F. Poggenhans, T. Kühner, M. Lauer, and C. Stiller, "Accurate and Efficient Self-Localization on Roads using Basic Geometric Primitives", in *2019 Int. Conf. Robotics and Automation (ICRA)*, 2019, pp. 5965–5971.

[11] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3D localisation", en, *Machine Vision and Applications*, vol. 25, no. 3, pp. 633–647, 2014.

[12] B. Riveiro, L. Díaz-Vilariño, B. Conde-Carnero, M. Soilán, and P. Arias, "Automatic Segmentation and Shape-Based Classification of Retro-Reflective Traffic Signs from Mobile LiDAR Data", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 1, pp. 295–303, 2016.

[13] H. Novais and A. R. Fernandes, "Community based repository for georeferenced traffic signs", in *2017 24º Encontro Português de Computação Gráfica e Interação (EPCGI)*, 2017, pp. 1–8.

[14] Z. Cui, Y. Liu, and F. Ren, "Homography-based traffic sign localisation and pose estimation from image sequence", *IET Image Processing*, vol. 13, no. 14, pp. 2829–2839, 2019.

[15] C. You, C. Wen, C. Wang, J. Li, and A. Habib, "Joint 2-d–3-d Traffic Sign Landmark Data Set for Geo-Localization Using Mobile Laser Scanning Data", *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 7, pp. 2550–2565, 2019.

[16] J. Balado, E. González, P. Arias, and D. Castro, "Novel Approach to Automatic Traffic Sign Inventory Based on Mobile Mapping System Data and Deep Learning", *Remote Sensing*, vol. 12, no. 3, p. 442, 2020.

[17] B. Soheilian, N. Paparoditis, and B. Vallet, "Detection and 3D reconstruction of traffic signs from multiple view color images", en, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 77, pp. 1–20, 2013.

[18] H. S. Lee and K. Kim, "Simultaneous Traffic Sign Detection and Boundary Estimation Using Convolutional Neural Network", *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1652–1663, 2018, Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[19] C. Liu, S. Li, F. Chang, and Y. Wang, "Machine Vision Based Traffic Sign Detection Methods: Review, Analyses and Perspectives", *IEEE Access*, vol. 7, pp. 86 578–86 596, 2019.

[20] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art", English, *Foundations and Trends in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020, Publisher: Now Publishers, Inc.

[21] J. Levinson, J. Askeland, J. Dolson, and S. Thrun, "Traffic light mapping, localization, and state detection for autonomous vehicles", in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 5784–5791.

[22] N. Fairfield and C. Urmson, "Traffic light mapping and detection", in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 5421–5426.

[23] G. Trehard, E. Pollard, B. Bradai, and F. Nashashibi, "Tracking both pose and status of a traffic light via an Interacting Multiple Model filter", in *17th International Conference on Information Fusion (FUSION)*, 2014, pp. 1–7.

[24] M. Diaz-Cabrera, P. Cerri, and P. Medici, "Robust real-time traffic light detection and distance estimation using a single camera", *Expert Systems with Applications*, vol. 42, no. 8, pp. 3911–3923, 2015.

[25] S. Hosseinyalamdary and A. Yilmaz, "A Bayesian approach to traffic light detection and mapping", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 125, pp. 184–192, 2017.

[26] M. B. Jensen, M. P. Philipsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi, "Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives", *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1800–1815, 2016.

[27] B. Mu, S.-Y. Liu, L. Paull, J. Leonard, and J. P. How, "SLAM with objects using a nonparametric pose graph", in *2016 IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, 2016, pp. 4602–4609.

[28] K. Ok, K. Liu, K. Frey, J. P. How, and N. Roy, "Robust Object-based SLAM for High-speed Autonomous Navigation", in *2019 Int. Conf. Robotics and Automation (ICRA)*, 2019, pp. 669–675.

[29] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Dual Quadrics From Object Detections as Landmarks in Object-Oriented SLAM", *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2019.

[30] S. Yang and S. Scherer, "CubeSLAM: Monocular 3-d Object SLAM", *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 925–938, 2019.

[31] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid, "Meaningful maps with object-oriented semantic mapping", in *2017 IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, 2017, pp. 5079–5085.

[32] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder, "Seamless scene segmentation", in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 8277–8286.

[33] M. Sons and C. Stiller, "Efficient Multi-Drive Map Optimization towards Life-long Localization using Surround View", in *2018 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, 2018, pp. 2671–2677.

[34] M. Schönbein and A. Geiger, "Omnidirectional 3D reconstruction in augmented Manhattan worlds", in *2014 IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2014, pp. 716–723.

[35] T. Strauss, "Kalibrierung von Multi-Kamera-Systemen - Kombinierte Schätzung von intrinsischem Abbildungsverhalten der einzelnen Kameras und deren relativer Lage zueinander ohne Erfordernis sich überlappender Sichtbereiche", PhD thesis, Faculty of Mechanical Engineering, Karlsruhe Institute of Technology, 2015.

[36] J. Beck and C. Stiller, "Generalized b-spline Camera Model", in *2018 IEEE Intell. Vehicles Symposium (IV)*, 2018, pp. 2137–2142.

[37] J. Kümmerle, T. Kühner, and M. Lauer, "Automatic Calibration of Multiple Cameras and Depth Sensors with a Spherical Target", in *2018 IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, 2018, pp. 1–8.

[38] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", in *KDD*, 1996, pp. 226–231.

[39] H. W. Kuhn, "The Hungarian method for the assignment problem", en, *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[40] J. Munkres, "Algorithms for the Assignment and Transportation Problems", *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[41] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset", *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[42] D. Gossow, A. Leeper, D. Hershberger, and M. Ciocarlie, "Interactive markers: 3-d user interfaces for ros applications [ros topics]", *IEEE Robotics Automation Magazine*, vol. 18, no. 4, pp. 14–15, 2011.

[43] J.-H. Pauls, K. Petek, F. Poggenhans, and C. Stiller, "Monocular localization in hd maps by combining semantic segmentation and distance transform", in *2020 IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, 2020, pp. 4595–4601.