Robust Monocular Localization in Sparse HD Maps Leveraging Multi-Task Uncertainty Estimation

Kürsat Petek*, Kshitij Sirohi*, Daniel Büscher and Wolfram Burgard

***Abstract*— Robust localization in dense urban scenarios using a low-cost sensor setup and sparse HD maps is highly relevant for the current advances in autonomous driving, but remains a challenging topic in research. We present a novel monocular localization approach based on a sliding-window pose graph that leverages predicted uncertainties for increased precision and robustness against challenging scenarios and per-frame failures. To this end, we propose an efficient multi-task uncertainty-aware perception module, which covers semantic segmentation, as well as bounding box detection, to enable the localization of vehicles in sparse maps, containing only lane borders and traffic lights. Further, we design differentiable cost maps that are directly generated from the estimated uncertainties. This opens up the possibility to minimize the reprojection loss of amorphous map elements in an association-free and uncertainty-aware manner. Extensive evaluation on the Lyft 5 dataset shows that, despite the sparsity of the map, our approach enables robust and accurate 6D localization in challenging urban scenarios using only monocular camera images and vehicle odometry.**

I. Introduction

Despite recent developments in the field of autonomous driving, HD maps remain an indispensable component in modern systems as they provide detailed information on the road infrastructure enabling various applications such as motion planning and enhanced perception. However, to utilize HD maps, it is necessary to accurately determine the vehicle pose within the map. To solve this localization task, high precision systems generally employ RTK-GNSS systems with integrated IMUs or vehicle odometry in a probabilistic fusion scheme [1], [2]. However, the high cost of equipment and limitations in dense urban environments render these methods limited to research and data generation. Therefore, most deployed autonomous vehicles utilize an HD map that stores information about the environmental elements like lane topology, traffic signs, and traffic lights to localize. With the availability of this information, landmark-based localization has gained interest. Such methods consist of a perception module to extract the necessary features from the sensor readings and continuously match them to the available map elements to constrain the vehicle pose [3], [4].

Supported by the extensive research and advances in deep learning, the perception modules in modern systems typically consist of a convolutional neural network (CNN) to extract the necessary features from the environment [5], [6]. Although state-of-the-art CNN architectures can provide a holistic understanding of the environment utilizing sensors such as cameras [7], [8] and LiDARs [9], most CNNs are not capable of providing reliable uncertainty estimate related to their predictions. The softmax operation is often employed, which overestimates the predictive probability of a network. This can compromise a localization algorithm's robustness and accuracy.

The task of uncertainty estimation with deep learning extends the standard neural network-based methods to additionally predict the associated uncertainty or confidence in the prediction. Popular uncertainty estimation methods primarily utilize the sampling-based methods [10], which are computation and time-intensive. While the research for sampling-free methods is gaining interest, current approaches focus on predicting the uncertainties for a single task like classification [11] or regression [12]. In contrast, an overall perception system of autonomous vehicles consists of various tasks, like segmentation and detection.

In this paper, we aim to solve the localization problem in challenging urban scenarios with a low-cost sensor setup and extremely sparse HD maps containing only lane borders and traffic lights. We present a novel monocular camera-based localization system that leverages the uncertainty estimations of our proposed multi-task perception module. Our novel perception module simultaneously predicts the uncertainties associated with semantics of the lane and with bounding box parameters of the traffic lights in a single pass.

The main contributions of this paper are: 1) a novel pose graph localization system robust in challenging scenarios by exploiting predicted uncertainties, 2) a multi-task uncertainty-aware perception module capable of simultaneously predicting semantic and regression uncertainties in a sampling free fashion, 3) a novel association-free and differentiable cost map generation module guided by prediction uncertainties.

We demonstrate the performance gain by incorporating uncertainties in our localization method by evaluating on the challenging Lyft 5 dataset [13].

II. Related Work

A. Localization

Recent works on localization mainly differ between the utilized map elements, the employed sensor setup, and the perception module. GNSS-based approaches fuse short-term accurate proprioceptive sensor information, e.g. vehicle

*These authors contributed equally. All authors are with the Department of Computer Science, University of Freiburg, Germany. This work was partly financed by the Baden-Württemberg Stiftung gGmbH (perception module) and by the Bundesministerium für Bildung und Forschung (localization method).

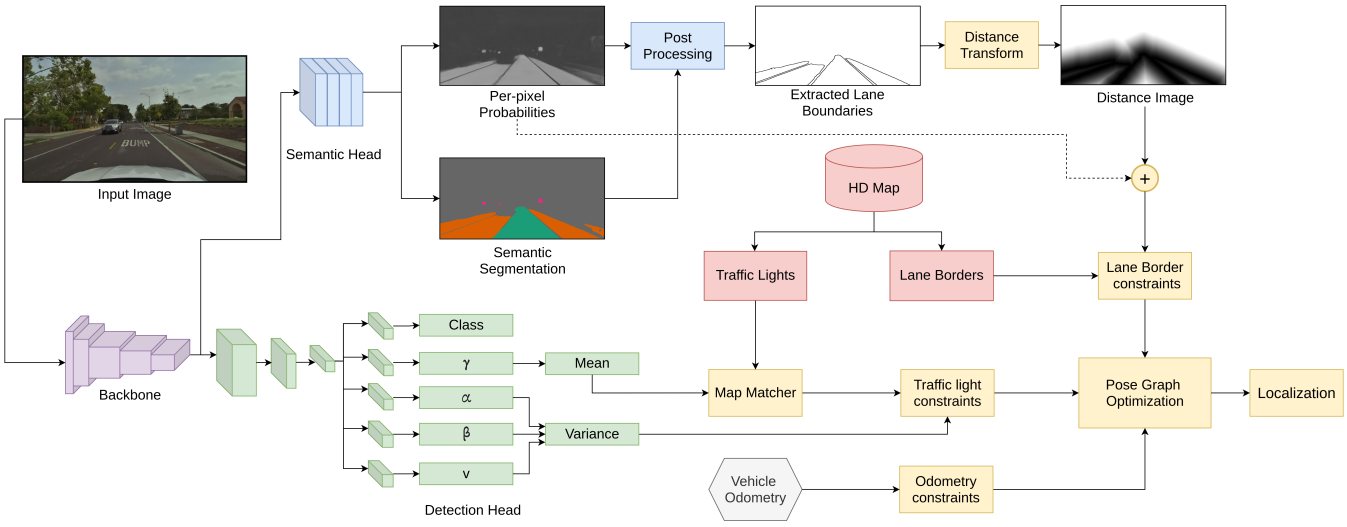

Fig. 1. Overview of our method. The input image is fed into a multi-head uncertainty-aware network with separate semantic (blue) and detector heads (green) to predict semantic and bounding box uncertainties together with their respective tasks. The predicted semantic probabilities, derived from uncertainties, are used to extract lane borders in the post-processing step. A distance transform is applied to these boundaries, followed by weighting with semantic probabilities to create a cost map. Traffic lights are matched to their corresponding detections by the map matcher. Finally, both perception constraints are set up and fed into the pose graph optimization along with odometry constraints for robust localization.

odometry or IMU, with long-term accurate GNSS information in a tightly coupled manner [1], [2]. However, the limitations of GNSS systems in dense urban scenarios affect the reliability of these methods. Other methods employ LiDAR maps for accurate localization [3], [4], but the reliance on the costly sensor or memory-intensive dense LiDAR maps impacts the scalability of such methods. Other approaches include specialized methods for extracting landmarks such as poles [14], [15], lane markings [5], [2], and facades [15]. Although being accurate, these methods require specialized detectors and mapping procedures for reliable localization.

Hence, deep learning-based methods have gained importance to render the localization flexible and scalable with the availability of additional data. Radwan et al. [16] propose to use a fully learning-based visual localization method that predicts the pose difference between consecutive images and the global pose of each frame. Pauls et al. [6] introduces a hybrid monocular localization method combining the advantages of deep learning and classical approaches. However, they use a pre-implemented network without adaptations according to the needs of localization tasks and the method is unaware of the inherent network uncertainties, rendering it unreliable in challenging environments.

B. Uncertainty Estimation

Uncertainties are classified into aleatoric (data) uncertainty to quantify the noise in data and epistemic (model) uncertainty to quantify uncertainty in model prediction due to lack of training data or insufficient knowledge of the model [17]. While it is possible to derive data uncertainty from data statistics or learn it with a network, it is harder to predict epistemic uncertainty due to the intractability of exact Bayesian inference for neural networks. To this end, most methods employ the popular sampling-based Monte Carlo (MC) dropout technique [10] and Bayesian neural networks (BNNs). For example, methods such as [18], [19], and [20], [21] employ modified versions of MC dropout to predict per pixel semantic and bounding box regression uncertainties, respectively.

The sampling-based approaches require multiple passes through a network or the predictions from multiple networks, rendering such approaches not fit for real-time applications. On the other hand, the sampling-free methods focus on predicting uncertainties in a single pass. In bounding box uncertainty estimation, approaches such as [22], [23] generally predict the aleatoric uncertainty for each of the bounding box parameters using a modified loss function by taking an extra variance term into account.

Sensoy et al. [11] introduce a sampling-free method called deep evidential learning to quantify the classification uncertainty by making the network collect evidence to predict higher-order prior distribution parameters. In this context, evidence denotes the magnitude of support the network predicts in favor of classifying a sample to a particular class. Capellier et al. [24] utilizes evidential deep learning to filter the object detections based on the uncertainty estimate for object classification in LiDAR point clouds. The approach in [12] proposes evidential deep learning for the task of monocular depth estimation, which is a regression task. Liu et al. [25] proposes to regress uncertainties related to control estimation in autonomous driving using evidential deep learning. These approaches show the strength of evidential deep learning by providing comparable or superior results to most sampling-based methods. Hence we utilize evidential deep learning to simultaneously predict semantic segmentation and bounding box detection uncertainties in a single pass.

III. Technical Approach

Our localization method consists of an uncertainty-aware perception module, a differentiable cost-map generator, a map matcher and a pose graph optimization module (see Fig. 1). The perception module incorporates a semantic

head for spatially unconstrained map elements, i.e. driveable areas, and a bounding box detection head for map elements with a finite extent like traffic lights.

The segmentation outputs are processed along with the estimated uncertainties to create a differentiable cost-map in the image plane which, in combination with the corresponding map elements, provides the lateral constraints for the camera pose from lane borders. The detection head detects traffic lights represented as bounding boxes together with the uncertainties associated with each parameter.

In the next step, the map matcher associates each potentially visible traffic light from the map with its counterpart from the detection module. The bounding boxes and the reprojections of potentially visible traffic lights serve as the inputs to compute the cost term. Traffic light constraints are set up based on these associations to penalize the point-to-point pixel distance between the instances and the reprojected traffic lights.

In the final step, the sliding-window pose graph optimization problem combines the constraints from the traffic lights and lane borders with odometry constraints to robustify the method and overcome per-frame failures in the perception module. In the end, optimization provides the most recent pose $\boldsymbol{p}^*$ as the localization result.

A. Perception Module

We use a convolutional neural network for the perception module. The architecture consists of a shared EfficientNet-B3 [26] backbone with a feature pyramid network on top [27]. The backbone learns features at multiple scales, utilized by separate semantic and detection heads.

1) Uncertainty-Aware Semantic Segmentation Head: Our semantic segmentation head is a modified version of the semantic head proposed in [7], which takes features at multiple scales and upscales them to a common scale followed by concatenation. We modify the semantic head by replacing the softmax at the end of the network with ReLU, which serves as an evidence signal of the model.

The evidential deep learning method proposes to estimate high order conjugate priors over the network output distribution to estimate the classification uncertainties [11]. We use the Dirichlet distribution as the prior for multinomial classification prediction per pixel, which is parameterized by N parameters $\alpha = [\alpha_1, .., \alpha_N]$ and the network is trained to predict α_i for each class i of total N classes. For semantic segmentation, the network predicts α for every pixel of the image.

We utilize the sum of squares form of the loss $L(\zeta)_i$ to penalize the misclassified pixels and the Kullback-Leibler (KL) divergence loss $\mathcal{L}_i^{\text{KL}}$ to predict high uncertainties for low evidence predictions for pixel i, as described by [11]. Moreover, as our application is semantic segmentation, we formulate the overall semantic loss as

$$\mathcal{L}_{sem} = \sum_{w=1}^{W}\sum_{h=1}^{H} L(\zeta)_{w,h} + \lambda_s \sum_{w=1}^{W}\sum_{h=1}^{H} L_{w,h}^{KL}, \qquad (1)$$

where W and H are the width and height of the image, respectively, and λ_s is the annealing coefficient. We use $\lambda_s = \min(1.0, t/4)$, where t is the ratio of the current iteration number and the total iterations per epoch.

2) Uncertainty-Aware Object Detection Head: For the detection head, we use a modified Faster-RCNN network to predict the class, bounding box parameters, and additional three parameters required for uncertainty estimation. We define the bounding box by the parameters (x_{min}, y_{min}, x_{max} and y_{max}). Hence, the estimation of the bounding boxes is formulated as a regression problem. The aim is to estimate the mean μ and the associated variance σ^2 for each of the four bounding box parameters.

As proposed by [12], we utilize Normal-Inverse-Gamma (NIG) distribution as conjugate prior for evidential regression learning. NIG is defined by four parameters, γ, α, β and υ. To this end, we extend the Faster-RCNN [28] network to have four separate branches to predict these parameters, as depicted in Fig. 1. In this formulation the mean value μ is given by γ. The associated aleatoric uncertainty is calculated as $U_{\text{a}} = \beta/(\alpha - 1)$ and the epistemic uncertainty as $U_{\text{e}} = U_{\text{a}}/\upsilon$.

We train our object detection head with the negative log-likelihood loss $\mathcal{L}_{\text{NLL}}$ for maximizing the model evidence or correct predictions and a regularizer term for penalizing errors scaled by the evidence $\mathcal{L}_{\text{R}}$ [12]. With the addition of the common objectness score loss $\mathcal{L}_{\text{os}}$ and the object proposal loss $\mathcal{L}_{\text{op}}$ [28], the total detection loss is defined as:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{NLL}} + \lambda_{\text{det}}\mathcal{L}_{\text{R}} + \mathcal{L}_{\text{os}} + \mathcal{L}_{\text{op}} \qquad (2)$$

The authors of [12] suggest using the scaling factor $\lambda_{\text{det}} = 0.01$ for the task of depth regression. However, the network tends to predict overconfident estimations with this value for the task of bounding box regression. Thus we have used $\lambda_{\text{det}} = 0.04$.

Finally, the overall loss is defined as $\mathcal{L} = \mathcal{L}_{sem} + \lambda\mathcal{L}_{\text{det}}$, where we use a scaling factor of $\lambda = 15$, since $\mathcal{L}_{sem}$ and $\mathcal{L}_{\text{det}}$ have different scales.

B. Differentiable Cost Map Generation

The availability of lane topologies for many roads in HD maps makes their usage highly relevant for the task of localization in autonomous driving [29]. This task is composed of detecting the lane borders and matching them to their counterparts in the HD map. As such, we want to detect any consistent longitudinal feature on the road as lane borders, such as lane markings, road boundaries, and parking zones.

The semantic head of our network provides segmentation for direct drivable and alternative drivable areas on the road with associated uncertainty values. Due to our well-calibrated uncertainty predictions (see Section IV-B) the probability map P_s consistently yields high uncertainty values for continuous longitudinal distortions on and off the road. Thus, for extracting the aforementioned longitudinal features, we simply threshold the uncertainty map I_{unc} with Otsu's method that maximizes the separability of the uncertainty

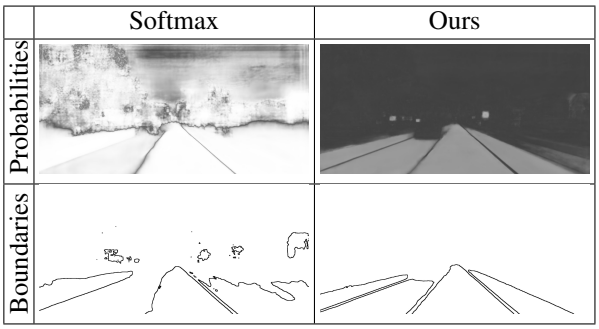

Fig. 2. Comparison of the predicted probabilities (upper) and borders extracted using probabilities (lower) between a simple softmax (left) and our uncertainty-aware semantic segmentation (right). Light areas represent high and dark regions represent low probability. The softmax tends to overestimate the probabilities for all regions, whereas our calibrated probability estimation assigns low values to non-lane areas and leads to better border estimations.

values [30] and mask out all non-lane classes to obtain the fully segmented lane borders. The result is an image I_{lb} containing only the lane borders depicted by boundaries in Fig. 2.

We optimize an error metric to constrain the vehicle pose to penalize the mismatch between the extracted lane borders and the lane topology map, reprojected onto the image. However, any direct association between the detected lane border pixels and reprojected map elements will be imperfect, since lane borders do not constrain in the longitudinal direction. To overcome this challenge, we apply a distance transform on the detected lane borders as proposed in [6]. The distance transform yields a cost-map $\boldsymbol{C}_s$ with each pixel containing the euclidean distance to the closest lane border in pixel space.

In the final step, we overlay $\boldsymbol{C}_s$ with the weighted probability map and apply a bi-cubic interpolation to obtain the final differentiable cost map $\boldsymbol{C}_{\mathrm{unc}}$. The probabilities show a smooth transition from the lane segments towards the approximate centerline of segmented lane borders yielding nonzero gradients that is beneficial for optimization. This allows for considering the uncertainties of the perception module throughout the whole extent of the lane borders.

C. Map Matching

The map matching step aims to associate the potentially visible set of reprojected map elements $\boldsymbol{X}^{\mathrm{tl}}$, the traffic lights, to the detections provided by the detection head of our perception module. First, we compute the per-pixel distance between each reprojected traffic light center and each detected bounding box center. This distance serves as a quality measure for each potential association. Second, due to the reliability of the map, the reprojected traffic lights are associated with the closest detections. Thus, even an erroneous pose yields correct associations as minor errors have almost no impact on the position of reprojections in the image plane for distant regions. Our perception module is capable of detecting traffic lights from a far distance of approximately 50m. Hence, we apply the map matching as early as possible and keep the correct initial associations until the vehicle moves past the traffic lights under consideration.

D. Sliding-Window Pose Graph Optimization

Due to the missing redundancy of map elements and the potential per-frame failures in the perception module in highly challenging scenarios, we choose to design a robust sliding-window pose graph optimization method [31]. This method optimizes N poses simultaneously, constrained by the detected features and the corresponding map elements. In order to obtain the final state vector $\boldsymbol{p}^*$, we optimize the cost function $J = J^{\mathrm{o}} + J^{\mathrm{lb}} + J^{\mathrm{tl}}$, accounting for the lane borders (lb), traffic lights (tl) and the odometry (o):

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \sum_{i \in \{\mathrm{lb,tl,o}\}} \sum_{k=1}^{N} J^i \left(\boldsymbol{p}_k, \boldsymbol{z}_k^i, \boldsymbol{m} \right), \tag{3}$$

where $\boldsymbol{z}_k^i$ are the detections of the measurement class i for pose $\boldsymbol{p}_k$ and $\boldsymbol{m}$ is the semantic HD map. This cost function can be further split into its error terms $\boldsymbol{e}_k^i$ and the corresponding information matrix $\boldsymbol{\Omega}_k^i$.

$$\begin{aligned} \boldsymbol{p}^* = & \arg\min_{\boldsymbol{p}} \sum_{i \in \{lb,tl\}} \sum_{k=1}^{N} \rho \left(\boldsymbol{e}_k^{\mathrm{i,T}} \left(\boldsymbol{p}_k \right) \boldsymbol{\Omega}_{\boldsymbol{k}}^{\boldsymbol{i}} \boldsymbol{e}_k^i \left(\boldsymbol{p}_k \right) \right) \\ & + \sum_{k=1}^{N-1} \left(\boldsymbol{e}_k^{\mathrm{o,T}} \left(\boldsymbol{p}_k, \boldsymbol{p}_{k+1} \right) \Omega_k^{\mathrm{o}} \boldsymbol{e}_k^{\mathrm{o}} \left(\boldsymbol{p}_k, \boldsymbol{p}_{k+1} \right) \right), \end{aligned} \tag{4}$$

where $\rho(x) = \log(1+x)$ denotes the Cauchy function, which robustifies the method by remapping the loss values via a logarithmic projection and effectively lowers the impact of outliers in the optimization.

The error term related to the lane borders, $\boldsymbol{e}_k^{\mathrm{lb}}$, is obtained by reprojecting the potentially visible map elements directly into the uncertainty cost map $\boldsymbol{C}_{\mathrm{unc}}$, using the forward pinhole camera model f_{cam} and the pose $\boldsymbol{p}_k$:

$$\boldsymbol{e}_k^{\mathrm{lb}} = \boldsymbol{C}_{\mathrm{unc}} \left(f_{\mathrm{cam}} \left(\boldsymbol{p}_k^{-1} \boldsymbol{X}^{\mathrm{lb}} \right) \right), \tag{5}$$

where $\boldsymbol{X}^{\mathrm{lb}}$ denotes the set of all lane border point positions under consideration.

The second error term, $\boldsymbol{e}_k^{\mathrm{tl}}$, directly penalizes the pixel distance between the bounding box and the associated traffic light points:

$$\boldsymbol{e}_k^{\mathrm{tl}} = \boldsymbol{X}_k^{\mathrm{bb}} - f_{\mathrm{cam}} \left(\boldsymbol{p}_k^{-1} \boldsymbol{X}^{\mathrm{tl}} \right), \tag{6}$$

where $\boldsymbol{X}^{\mathrm{bb}}$ denotes the pixel positions of the set of traffic lights detected in frame k represented as bounding boxes (bb). Due to the finite extents of traffic lights, a direct association is possible and no additional cost map has to be generated. This error term is differentiable and can directly be incorporated into the optimization problem. The probabilities estimated by our uncertainty-aware perception module are considered in the information matrix.

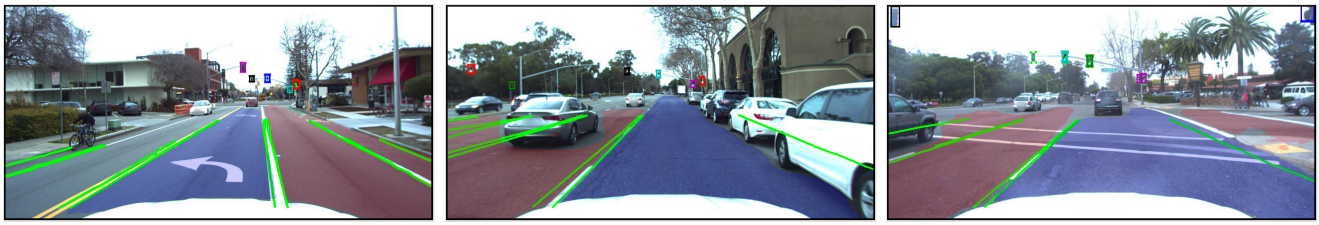

Fig. 3. Reprojected map elements based on our localization results in three different scenarios. Despite occlusions and challenging intersections, our method yields accurate localization results even with a small set of lane borders or traffic lights. The semantics of lanes and bounding boxes are predicted by the perception network. The width of bounding boxes represents the variance predicted for the corresponding edge (best viewed at x4 zoom scale).

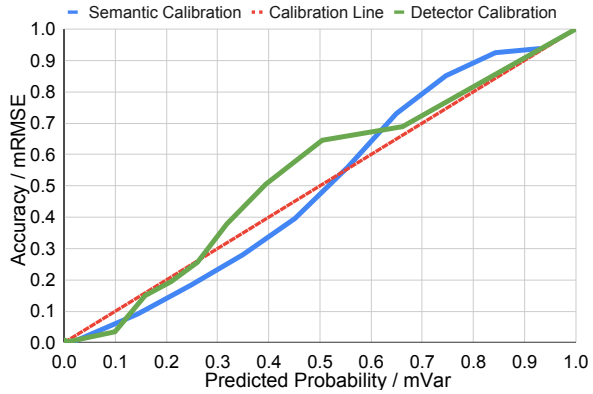

Fig. 4. Calibration plot for the estimated semantic and detection uncertainties. For semantic segmentation, it shows the accuracy vs. the predicted probability (blue), and for detection, it shows the mRMSE vs. the mVar metric (green). Both uncertainty estimations follow the calibration line (red) in close vicinity, signifying well-calibrated uncertainty estimations.

Finally, any displacement on consecutive poses $\boldsymbol{p}_k$ and $\boldsymbol{p}_{k+1}$ imposed by the optimizer that deviates from the odometry $\boldsymbol{\Delta}_{k\to k+1}$ is penalized by

$$\boldsymbol{e}_k^{\circ}\left(\boldsymbol{p}_k, \boldsymbol{p}_{k+1}\right) = \boldsymbol{p}_k^{-1}\boldsymbol{p}_{k+1} - \boldsymbol{\Delta}_{k\to k+1}^{\text{meas}}. \tag{7}$$

With the final cost function being set up, we can now localize the vehicle within the map by optimizing this overall cost. Alternatively, this method can be executed in a single frame setting by dropping the odometry constraints. This, however, is only done for evaluation purposes (see Sec. IV-C).

IV. Experimental Evaluation

A. Dataset

For evaluating our localization system, we explore the Lyft5 [13] dataset, which provides 6D localization ground truth and a semantic HD map containing the lane topologies and the sparse instance elements, like traffic lights. Example images are given in Fig. 3. However, the scenes are only about 25 seconds long, and the odometry information is also missing. Thus, we order and stitch the scenes together using ground truth poses to create a long continuous sequence with a length of 2.6 km and in a highly populated and challenging urban area. We create the odometry by utilizing the ground truth poses to predict noisy odometry signals according to the velocity-based motion model. The accuracy obtained by the emulated odometry is equally or less precise than modern car odometry systems evident from the longitudinal drift presented in Fig. 5.

We train our perception network on the bdd100k dataset [32], containing 70,000 images for training and 10,000 images for validation. We train the semantic head to predict the direct drivable area and alternative driveable area as introduced in [32]. Similarly, the detection head is trained to predict bounding boxes for the traffic light class. To evaluate the performance of our uncertainty estimation, we utilize the validation set, which the network never used either during training or for uncertainty calibration. Please note that the network never utilized any data from the Lyft5 [13] dataset during training.

B. Uncertainty Estimation

We evaluate the performance of the semantic segmentation uncertainty estimation to make sure that they are well-calibrated. We report the values of the Expected Calibration Error (ECE). For calculating this metric, the predicted probability axis is divided into J equally spaced bins, and for each bin the average accuracy $\text{acc}(B_j)$ and average predicted confidence $\text{conf}(B_j)$ are computed. Then the ECE is given as

$$\text{ECE} = \sum_{j=1}^{J} \frac{n_j}{N} \left|\text{acc}(B_j) - \text{conf}(B_j)\right|, \tag{8}$$

where n_j is the number of samples in bin j and N is the total number of samples. The ECE value depicts the deviation from the optimal calibration line. We achieve an ECE of 5.3%. For bounding box detection, we utilize the Expected Normalized Calibration Error (ENCE) [33] which is similar to the ECE. As bounding box detection is a regression task, ENCE reflects the relation between the predicted variance and the Root Mean Square Error (RMSE). The detector achieves an ENCE value of 11.5% which is reasonable in comparison to the value of 8.5% for [34] reported by [33] on the KITTI dataset. Note that the lower value signifies better performance in both tasks.

In addition, we show the calibration plot of the accuracy vs. the predicted probability, for semantic segmentation and predicted mean-variance (mVar) per bin vs. (mRMSE) per bin [33] for bounding box detection, see Fig. 4. The desired result is a response that is close to the $y = x$ line. We observe that both tasks follow the calibration line in close vicinity and hence provide meaningful and usable uncertainties.

TABLE I

SINGLE IMAGE LOCALIZATION SUCCESS RATE (S.R.) AND MEAN TRANSLATIONAL AND ROTATIONAL ERRORS IN PERCENT, METERS OR DEGREES.

$\delta(m)$	$\delta(°)$		s.r.	lat	z	yaw	pitch	roll
± 0.5	± 2.5	D	52	0.26	0.33	1.17	1.27	1.19
		D+	78	0.24	0.23	1.05	1.00	0.97
± 0.75	± 5.0	D	49	0.26	0.42	1.21	1.61	1.63
		D+	68	0.24	0.33	1.04	1.38	1.25
± 1.0	± 7.5	D	45	0.25	0.51	1.21	2.19	2.25
		D+	57	0.24	0.41	1.00	1.80	1.41

C. Single Image Localization

To showcase how uncertainties help to improve the accuracy and reliability of the localization method, we evaluate our approach by using a single frame for localization. First, we add translational and a rotational noise, sampled from three different settings of a uniform distribution, onto the ground truth pose to obtain a distorted pose. Second, we initialize the localizer with this distorted pose and relocalize the camera within the lane using the lateral constraints. Here, we omit the odometry constraints to cancel out their impact on the localization result.

We compare two settings, one with the distance transform directly applied on the predicted lane classes without utilizing the uncertainties (D), and the other using the uncertainties for segmenting all lane/road borders (D+). The results are reported in Table I. In addition to the translational and rotational accuracy, we also report the localization success rate (s.r.). We define a success as a final lateral error below 0.5m and a yaw angle error below 2.5°, which is sufficient to initialize a localization system. Only successful cases contribute to accuracy evaluation. We do not report longitudinal pose errors, since this direction is not constrained by the lanes in single images.

Our uncertainty-based method outperforms the method operating directly on the semantic outputs for every single measure. While the lateral errors show comparable results, we observe that the additional redundancy given by the uncertainty-based method resolves ambiguities w.r.t. the height error and the rotational error. Our method also shows a much higher success rate, proving the robustness towards localization errors.

D. Pose Graph Localization

We evaluate our pose graph localization approach on the Lyft 5 dataset, which presents challenging urban scenarios with parking zones, occluded lane borders, and a large number of intersections, where lane borders are barely marked out, see Fig. 3. The mean lateral and yaw errors of 0.19 m and 1.3°, presented in Table II, show that our method can keep up with the state of the art, which yields errors in the range of 0.1 - 0.3 m and 1 - 2°, respectively [3], [35], [36]. However, these aforementioned methods either utilize a precise LiDAR sensor or memory-intensive maps.

Though our method is laterally accurate on average, complex intersections, and highly populated scenes can degrade the performance due to missing features at intersections and frequent occlusions through parked cars and heavy traffic, see Fig. 5 (top). However, our method is able to track the pose even when only a small subset of lane borders is visible.

TABLE II

POSE GRAPH LOCALIZATION RESULTS IN METERS OR DEGREES RELATIVE TO THE REFERENCE PROVIDED BY LYFT 5.

	lon	lat	z	yaw	pitch	roll
RMSE	0.61	0.25	0.28	1.30	0.59	0.63
MAE	0.45	0.19	0.17	1.01	0.45	0.42

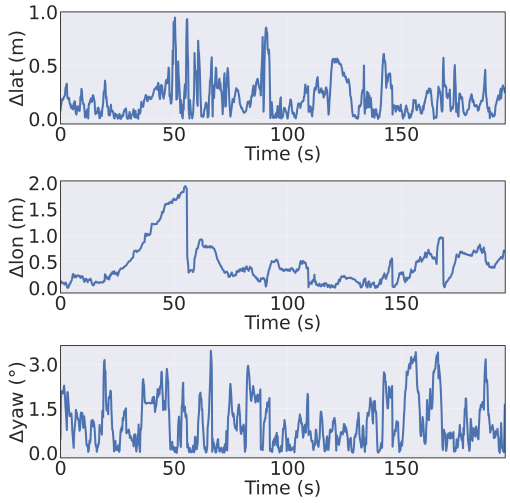

Fig. 5. Lateral, longitudinal and yaw errors for our pose graph localization throughout the whole test sequence.

In contrast, the longitudinal pose tends to drift in long driving sequences due to the sparsity of longitudinal cues, see Fig. 5 (middle). It is noticeable that even from large initial errors, the longitudinal pose converges very fast to feasible solutions as soon as traffic lights, or lane features that can break the symmetry in the longitudinal direction, appear. Though relying only on a small set of traffic lights for constraining the longitudinal pose, our localizer manages to keep the mean longitudinal error impressively small by only using constraints in the image plane.

V. CONCLUSION

In this work, we proposed a novel monocular localization system that incorporates predicted uncertainties into a pose graph optimization framework. The uncertainties help to attain robustness in challenging urban scenarios using only sparse map features. As a crucial part of our approach, we presented a novel multi-task uncertainty estimation method that demonstrated the capability to simultaneously learn meaningful uncertainties for semantic segmentation and object detection in a single pass. Even though we only use a camera and a sparse map, we demonstrate through our experiments that our approach performs on par with methods that utilize expensive sensor setups or dense maps.

REFERENCES

[1] K. Jo, K. Chu, and M. Sunwoo, "Interacting multiple model filter-based sensor fusion of gps with in-vehicle sensors for real-time vehicle positioning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 329–343, 2011.

[2] M. Schreiber, H. Königshof, A.-M. Hellmund, and C. Stiller, "Vehicle localization with tightly coupled gnss and visual odometry," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 858–863.

[3] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d lidar maps," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1926–1931.

[4] D. Cattaneo, M. Vaghi, A. L. Ballardini, S. Fontana, D. G. Sorrenti, and W. Burgard, "Cmrnet: Camera to lidar-map registration," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1283–1289.

[5] F. Poggenhans, N. O. Salscheider, and C. Stiller, "Precise localization in high-definition road maps for urban regions," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 2167–2174.

[6] 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 International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4595–4601.

[7] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder, "Seamless scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8277–8286.

[8] R. Mohan and A. Valada, "Efficientps: Efficient panoptic segmentation," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1551–1579, 2021.

[9] K. Sirohi, R. Mohan, D. Büscher, W. Burgard, and A. Valada, "Efficientlps: Efficient lidar panoptic segmentation," *arXiv preprint arXiv:2102.08009*, 2021.

[10] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[11] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *arXiv preprint arXiv:1806.01768*, 2018.

[12] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *arXiv preprint arXiv:1910.02600*, 2019.

[13] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska *et al.*, "Lyft level 5 av dataset 2019," *urlhttps://level5. lyft. com/dataset*, 2019.

[14] A. Schaefer, D. Büscher, 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)*. IEEE, 2019, pp. 1–7.

[15] 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 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5965–5971.

[16] N. Radwan, A. Valada, and W. Burgard, "Vlocnet++: Deep multitask learning for semantic visual localization and odometry," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4407–4414, 2018.

[17] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.

[18] J. Mukhoti and Y. Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *arXiv preprint arXiv:1811.12709*, 2018.

[19] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun, "Efficient uncertainty estimation for semantic segmentation in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.

[20] A. Harakeh, M. Smart, and S. L. Waslander, "Bayesod: A bayesian approach for uncertainty estimation in deep object detectors," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 87–93.

[21] F. Kraus and K. Dietmayer, "Uncertainty estimation in one-stage object detection," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 53–60.

[22] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2888–2897.

[23] M. T. Le, F. Diehl, T. Brunner, and A. Knol, "Uncertainty estimation for deep neural object detectors in safety-critical applications," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3873–3878.

[24] E. Capellier, F. Davoine, V. Cherfaoui, and Y. Li, "Evidential deep learning for arbitrary lidar object classification in the context of autonomous driving," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1304–1311.

[25] Z. Liu, A. Amini, S. Zhu, S. Karaman, S. Han, and D. L. Rus, "Efficient and robust lidar-based end-to-end navigation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 247–13 254.

[26] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[29] N. Homayounfar, W.-C. Ma, J. Liang, X. Wu, J. Fan, and R. Urtasun, "Dagmapper: Learning to map by discovering lane topology," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2911–2920.

[30] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[31] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based slam," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 4, pp. 31–43, 2010.

[32] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[33] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, "Evaluating and calibrating uncertainty prediction in regression tasks," *arXiv preprint arXiv:1905.11659*, 2019.

[34] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2796–2804.

[35] Y. Zhang, L. Wang, X. Jiang, Y. Zeng, and Y. Dai, "An efficient lidar-based localization method for self-driving cars in dynamic environments," *Robotica*, pp. 1–18, 2021.

[36] H. Yin, L. Tang, X. Ding, Y. Wang, and R. Xiong, "Locnet: Global localization in 3d point clouds for mobile vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 728–733.