IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 April 2023, accepted 6 June 2023, date of publication 14 June 2023, date of current version 20 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3286310

RESEARCH ARTICLE

Toward Precise Ambiguity-Aware Cross-Modality Global Self-Localization

NIKLAS STANNARTZ[1,*], STEFAN SCHÜTTE[1,*], MARKUS KUHN[2], AND TORSTEN BERTRAM[1]

[1]Institute of Control Theory and Systems Engineering, TU Dortmund University, 44227 Dortmund, Germany
[2]ZF Automotive Germany GmbH, 40549 Düsseldorf, Germany
[*]Niklas Stannartz and Stefan Schütte contributed equally to this work.

Corresponding author: Niklas Stannartz (niklas.stannartz@tu-dortmund.de)

This work was supported by Deutsche Forschungsgemeinschaft and Technische Universität Dortmund/TU Dortmund University within the funding program Open Access Costs.

ABSTRACT There are significant advances in GNSS-free cross-modality self-localization of self-driving vehicles. Recent methods focus on learnable features for both cross-modal global localization via place recognition (PR) and local pose tracking, however they lack means of combining them in a complete localization pipeline. That is, a pose retrieved from PR has to be validated if it actually represents the true pose. Performing this validation without GNSS measurements makes the localization problem significantly more challenging. In this contribution, we propose a method to precisely localize the ego-vehicle in a high resolution map without GNSS prior. Furthermore, sensor and map data may be of different dimensions (2D / 3D) and modality, i.e. radar, lidar or aerial imagery. We initialize our system with multiple hypotheses retrieved from a PR method and infer the correct hypothesis over time. This multi-hypothesis approach is realized using a Gaussian sum filter which enables an efficient tracking of a low number of hypotheses and further facilitates the inference of our deep sensor-to-map matching network at arbitrarily distant regions simultaneously. We further propose a method to estimate the probability that none of the currently tracked hypotheses is correct. We achieve successful global localization in extensive experiments on the MulRan dataset, outperforming comparative methods even if none of the initial poses from PR was close to the true pose. Due to the flexibility of the approach, we can show state-of-the-art accuracy in lidar-to-aerial-imagery localization on a custom dataset using our pipeline with only minor modifications of the matching model.

INDEX TERMS Vehicle self-localization, cross-modality localization, global localization, place recognition, multi-hypothesis localization, HD map, automated driving.

I. INTRODUCTION

HD maps may support the automated vehicle in various functionalities like navigation, perception, trajectory planning and situation prediction [4]. To enable those functionalities, an accurate pose estimate of the ego-vehicle within the map is required through localization. The localization task may be split up into first solving the global localization problem, i.e. finding the (rough) initial vehicle pose, and subsequently keeping track of the pose while enhancing its accuracy over time, assuming that the current pose estimate is close to the true vehicle pose [5]. While common GNSS-based localization has the advantage of solving global localization, its application for precise pose tracking in HD maps is generally insufficient as its accuracy often lies in the range of meters even in combination with dead reckoning [6]. This problem can be exacerbated in urban regions where obstacles, such as tall buildings, block the direct GNSS signal path which may lead to large GNSS errors, e.g. due to multipath effects [7], or even complete outages. Moreover, recent studies reveal that GNSS is additionally prone to jamming or spoofing [8] which may be another cause for localization failure using GNSS.

To overcome the drawbacks of GNSS-based localization and enhance the localization accuracy and robustness, HD maps contain georeferenced metric information of the static environment that may be detected by the perception sensors of the automated vehicle, including camera, lidar, and

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

radar. Matching corresponding sensor and map data enables both GNSS-free global localization, e.g. through place recognition (PR) [9], [10], [11] and precise local pose tracking [12], [13]. Previous approaches employed the same sensor modality for the map creation and subsequent localization including approaches based on lidar [14], [15], [16], [17], camera [18], [19], [20], and radar [21], [22], [23] sensors as well as combinations of them [6], [24], [25]. However, using the same sensor modality for mapping and localization either requires a certain sensor to be available on the vehicle for facilitating localization or multiple sensor-specific HD maps of the same environment which increases data storage and the effort of map creation and maintenance. Moreover, it limits the availability and utility of such maps, as innovation in the sensor technology used in vehicles may introduce the necessity for frequent updates to the mapping vehicle fleet.

Therefore, cross-modality localization approaches, i.e. employing different sensor modalities for the map creation and subsequent localization, have potential to be a more cost-effective alternative as they drastically increase the utility of the created HD maps and increase reusability. HD maps may thus be provided by third-party mapping companies [26] as well as publicly available data, e.g. from aerial imagery [27].

There are different approaches to realize cross-modality in localization. One approach is to define handcrafted static features within the environment, also called landmarks, that may be detected by different sensor modalities. Common landmark types are lane [28], [29] and road [30] markings, poles [31], road signs [32] as well as generic low-level features [33]. These approaches however, require knowledge about the environment, the sensor used on the vehicle, and additional effort to identify reliable landmark types while discarding much of the available sensor and map data. Moreover, this usually limits the usabiltiy of those maps to areas where the predefined landmark types are present and relevant for localization.

Recent works thus focus on *learnable* features for localization. Compared to handcrafted methods, learning-based approaches are driven towards finding reliably detectable features that are useful for the localization task based on a learning strategy and are not limited to those a human annotator deems useful. Especially for cross-modality localization, the identification of reliable landmarks for various sensor and map modalities is non-trivial. Here, learning-based approaches have become the state-of-the-art for both cross-modal PR [1], [11], [34], [35], [36] as well as local pose tracking, achieving localization accuracies below 1 m for various sensor modalities, including radar-to-lidar [2], [37], [38], range-to-aerial-imagery [35], [39], [40], [41] and camera-to-aerial-imagery, also called cross-view geo-localization (CVGL) [27], [41], [42].

Most of these approaches estimate the current pose by correlating or registering learned representations of the sensor image with a map section around the current estimate. Hence, they require that a prior pose estimate *close* to the ground truth (GT) pose is given or provided by a global localizer like a PR algorithm. Here, the term ''close'' is defined by the search range of the employed correlation or registration method and may amount up to 20 m to 40 m [2], [27], [40], [42]. However, assuming that a prior pose near the true pose is given is not viable for real world applications when GNSS is not available, since current PR algorithms are imperfect [1], [35]. Therefore, when using a PR algorithm for initialization, it is initially unclear if the retrieved pose is close to the true pose at all. However, for safety critical applications like automated driving, localization systems need to be self-assessing and output a confidence measure of the current estimate [43], [44]. To tackle this problem we propose multiple extensions to previous works: inspired by [45], [46], and [47], we propose to initialize and track multiple hypotheses during initialization using a Gaussian sum filter (GSF) where every hypothesis is represented by a weighted Gaussian. Here, the weight corresponds to the estimated probability of the corresponding hypothesis representing the true pose. The hypotheses are initialized based on the top-n retrieval results of a PR algorithm. Subsequently, the hypotheses are tracked over time whereas their weights are determined based on how consistent the odometry measurements are with the observations at the estimated locations of the map. The observations are obtained from our proposed deep correlation-based sensor-to-map matching network that is not only capable of extracting meaningful features for localization from sensor and map data of different modalities, e.g. radar sensor and lidar map, but the data may also have different dimensions, e.g. registering a 2D radar sensor image to a 3D lidar point cloud map. We realize this by adding a short variant of Point Voxel Convolution (PVConv) [48] as input layers which realizes a learnable projection of three-dimensional input data to two-dimensional features, thus keeping the subsequent correlation efficient in the SE2 space. Fig. 1 visualizes our method for a global radar-to-lidar localization initialization experiment on the KAIST02 sequence of the MulRan dataset [3]. Note, that no GNSS is used for initialization, but only retrieved poses from a PR method [1], radar odometry estimates [2] and pose observations from our proposed matching model. Note, that the correct hypothesis (*framed red*) may be inferred over time although it was the least likely hypothesis at the time of initialization. Furthermore, the tracking of multiple hypotheses enables the recognition of ambiguous scenarios when multiple hypotheses keep a non-negligible probability over time. The localization system can report being unavailable in these situations to avoid the dissemination of false information leading to safety-critical localization failures. To account for the case that none of the top-n retrieved poses of the PR algorithm is correct, we incorporate the concept of tracking a *null hypothesis probability* [46], [47] which corresponds to the probability that all the currently tracked hypotheses are incorrect. This allows us to find the

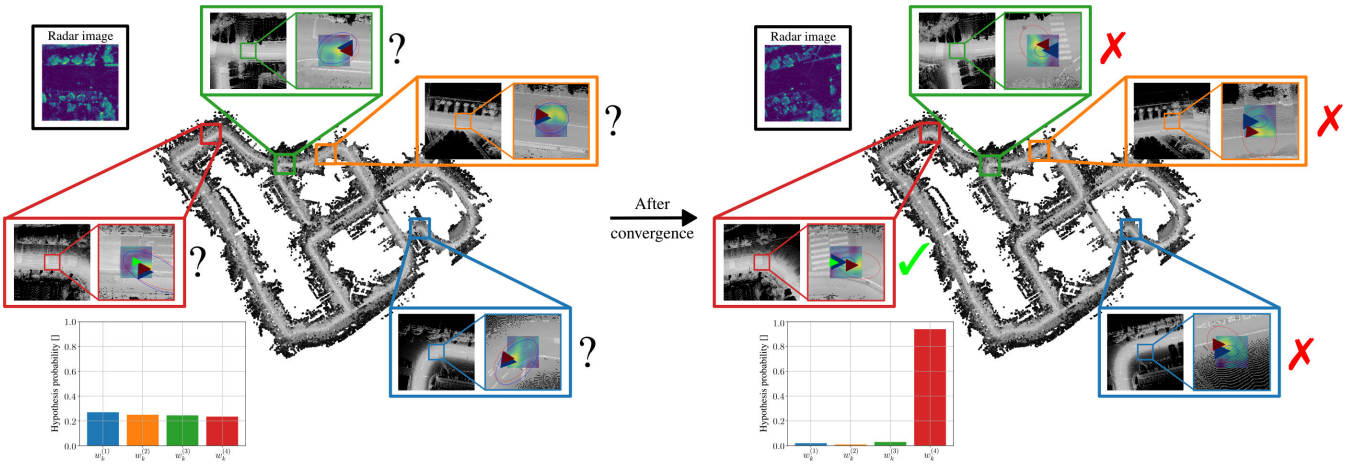


FIGURE 1. **Overview of our proposed ambiguity-aware cross-modality (in the depicted case radar-to-lidar) global self-localization approach. In this example, our method is initialized with four hypotheses, represented as a Gaussian mixture, based on the top-4 retrievals of a state-of-the-art radar-to-lidar place recognition method [1]. Subsequently, the correct, but initially least likely, hypothesis (marked in *red*) may be inferred over time using a Gaussian sum filter through the fusion of pose observations, provided by our correlation-based sensor-to-map matching model, and odometry estimates obtained from a given radar odometry [2]. The figure shows the situations at the time of initialization (left) and after convergence (right) from a radar-to-lidar global localization initialization experiment on the KAIST02 sequence of the MulRan dataset [3]. The bar plots depict the probabilities of the different hypotheses highlighted with the same colors in the zoomed in sections of the map. Within the zoomed in sections, the *blue triangle and covariance ellipse* denote an estimated hypothesis while the *red triangle and covariance ellipse* denote the pose observation resulting from a Gaussian approximation of the correlation output tensor of our matching model. Finally, the *green triangle* denotes the true pose.**

correct pose during initialization even if all initially retrieved hypothesis were wrong. While other approaches are also able to track multiple hypotheses, e.g. using a histogram filter (HF) [17] or a particle filter (PF) [37], [49], the GSF can more efficiently represent and maintain a multimodal posterior distribution using a low number of hypotheses while the hypotheses may be arbitrarily distant of each other. This would not be possible when using a HF for estimation as the computational demand would rapidly increase when increasing the spatial extent of the covered area for estimation due to the discretization of the state space. Similarly, a PF would require several thousands of particles for a reasonable performance [50], [51] which exceeds the computational resources available as well when provided as a batch to a deep correlation-based matching model. The GSF formulation therefore simultaneously enables us to significantly increase the search space of our matching model by providing the model a batch of map tiles as input based on the currently tracked hypotheses. The contributions of our work are therefore the following:

- We present a localization pipeline to solve both global localization and local pose tracking without any GNSS prior using only learned features of range sensor input and map data of different sensing modalities as well as dimensions by adding a stack of PVConv input layers to our deep correlation-based sensor-to-map matching model. Furthermore, through the employment of a multi-hypothesis localization approach, we provide means to validate which of the retrieved poses from a PR method is actually close to the true pose by determining the hypothesis that best explains the received odometry measurements and pose observations from our sensor-to-map matching model over time. Additionally, we integrate the concept of a null hypothesis proposed by [46] and [47] within our framework to simultaneously estimate the probability that none of the currently tracked poses is correct and show that we can successfully localize the ego-vehicle even if none of the initially retrieved poses were close to the true pose.
- We propose to implement the multi-hypothesis localization framework using a Gaussian sum filter (GSF) [45], [46], [47]. As opposed to histogram or particle filters, this formulation enables us to simultaneously evaluate our sensor-to-map matching network at different regions in the map that may be arbitrarily distant to each other while determining the most likely hypothesis over time through the fusion of odometry measurements and map-based pose observations while keeping the computational demand bounded. To account for misdetections and clutter measurements we employ the complete measurement likelihood formulation for single-object tracking (SOT) [52], [53], [54] which not only provides means for outlier rejection, but also augments the probability calculation of the hypotheses.
- We provide an empirical calibration of the proposed sensor-to-map matching model and present a method to determine meaningful values for the filter parameters prevalent in the measurement likelihood calculation based on a per-frame evaluation on the validation set.
- We show in extensive experiments that our method not only achieves state-of-the-art accuracy in radar-to-lidar and lidar-to-aerial-imagery localization, but also enables successful localization even if none of the initially retrieved poses from the PR was close to the true pose which has not been possible with previous works.

II. RELATED WORK

A. LEARNING-BASED CROSS-MODALITY POSE TRACKING

To achieve cross-modality pose tracking, current approaches first bring both the sensor and map data into a common representation and subsequently determine the pose offset between both representations to obtain an accurate localization estimate [2], [27], [35], [39], [40], [41], [55], [56]. To obtain an end-to-end learnable approach the operations within this pipeline are required to be differentiable with respect to the input data [17]. While a differentiable transformation into a common representation is easily realizable through CNN-based architectures [2], [17], [27], [35], [39], [40], [41], different methods exist to obtain a differentiable pose offset regression. Here, the most popular approaches perform differentiable pose offset regression through cross-correlation [27], [39], [40], [41], similarity-based offset estimation [2] or registration [35]. Cross-correlation based methods have the advantage that the correlation can be applied in the frequency domain and may be efficiently performed on a GPU, rendering the operation to be real-time capable [17], [57], [58]. Moreover, opposed to a registration-based method, the result of the cross-correlation corresponds to a probability distribution that may be further exploited as a measure of uncertainty [41], [56], [57]. Furthermore, state-of-the-art cross-modality correlation-based matching models achieve sub-meter level accuracy [27], [41]. In general, the presented pose tracking methods have in common that they require a global initial pose estimate close to the true pose. Here, the definition of ''close'' depends on the search range of the offset regression method ranging from 6 m [2] to 40 m [42] in lateral and longitudinal direction.

B. GNSS-FREE CROSS-MODALITY GLOBAL LOCALIZATION

When GNSS is unavailable, global localization becomes significantly more challenging as the global pose has to be solely inferred from perception, motion measurements and map data, especially when the perception and map data have different modalities. Current approaches tackle the problem either using a particle filter (PF) [59], [60] or a PR method [1], [35]. The PF approaches typically require knowledge about the location of the streets of the environment to uniformly initialize particles on every street. Yan et al. [59] obtain this information directly from OpenStreetMap (OSM) whereas Miller et al. [60] preliminarily perform a semantic segmentation on aerial imagery to segment the roads. Both approaches converge around the true pose with mean translation errors between 2 m to 20 m and average convergence times between 17 s to 75 s using measurements of 3D lidar sensor to update the particle weights. Opposed to PF approaches, PR-based global localization methods require a database of ''places'' typically represented as learned descriptors that capture the modality-specific appearance of each place [1], [35]. Current methods achieve top-1 recalls between 25 % to 65 % for radar-to-lidar [1] and 4 % to 15 % for lidar/radar-to-aerial-imagery [35] PR with metric accuracies around 3 m.

The above mentioned approaches enable successful cross-modal global localization without any GNSS prior, however, they differ in their efficiency and scalability. A PF requires a minimum initial particle density over the complete map to ensure successful convergence. Blanco-Claraco et al. [51] empirically determined a required particle density of $\approx 2\frac{\text{particles}}{\text{m}^2}$ for successful convergence in 3D lidar localization which requires $>$ 200 000 particles for an area of only 420 m $\times$ 320 m in size leading to a mean runtime of more than 2 s per time step when assuming an average execution time of 0.01 ms per particle. PR methods on the other hand may exploit efficient nearest neighbor queries achieving retrieval runtimes of around 1 ms for a corresponding database of similar size when assuming a density of 2 map images per m^2 [61]. With feature extraction times of around 10 ms, PR methods enable manageable runtimes for global localization.

A drawback when using PR for initialization it is preliminarily unknown which or if any of the retrieved PR poses is actually close to the true pose. Here, multi-hypothesis localization approaches provide means for detecting ambiguous scenarios and thus may be used for an integrity check of the localization estimate [62], [63], [64], [65], [66] as well as estimating the probability that none of the hypotheses is correct [46], [47].

C. AMBIGUITY-AWARE MULTI-HYPOTHESIS VEHICLE LOCALIZATION

In [67] and [68], the authors developed a multi-hypothesis localization approach to determine the vehicle's location in an OSM road network, solely based on motion estimates from a visual odometry method. The developed filtering algorithm was based on a Gaussian mixture representation of the posterior distribution of the state whereas the state variables were defined as the previous and current distance along a corresponding street segment as well as the previous and current orientation w.r.t. to the segment. Their localization system becomes available when there is a single mode in the posterior for at least 10 s. Both, Rabe et al. [62], [63] and Li et al. [64], [65], [66] employ a PF for an ambiguity-aware localization approach within a vectorized HD map. In [62] and [63], the localization system is only deemed available if the sum of the particle weights within a corresponding HD map segment exceeds a threshold. Similarly, the authors of [64], [65], and [66] require that after fault detection and the pruning of low-likely hypotheses only a single hypothesis, i.e. particle set on the same HD map lane segment, remains for the localization system to become available. Kim et al. [50] set a threshold on the standard deviation of the particles as a condition for the filter to be converged and thus the localization to be available. In [60], the convergence of a PF-based localization approach is automatically detected as well, despite the conditions for detecting convergence are not explicitly defined. We assume that the convergence criteria are similar to [50] since this approach is employed as a baseline method for comparison.

D. PLACEMENT OF OUR WORK

Our work relies on a deep correlation-based sensor-to-map matching model which allows for high localization accuracy while providing information of uncertainty of the pose estimate. Furthermore, to allow the processing of three-dimensional sensor and map data while keeping the correlation efficient the SE2 space, we add PVConv (Point-Voxel Convolution) layers [48] to the input layer which realizes a learnable projection of 3D points on the two-dimensional feature space. In contrast to previous works in the field of learning-based cross-modality pose tracking (cf. Sect. II-A), we simultaneously provide means for GNSS-free global localization by initializing our method with the top-n retrieval results of a PR algorithm and subsequently determine which of the retrieved poses is actually close to the true pose using a multi-hypothesis approach based on a Gaussian sum filter (GSF) [45]. This allows us to track a low number of hypotheses using the same precise matching model while we can evaluate our model simultaneously at arbitrarily distant regions by providing a small batch of map tiles centered around the currently tracked hypotheses. Compared to other GNSS-free cross-modal global localization methods (cf. Sect. II-B) we achieve state-of-the-art decimeter-level localization accuracy while our method is simultaneously more scalable to large maps than PF-based solutions. Moreover, through the introduction of a null hypothesis [46], [47], we additionally estimate the probability that none of the initially retrieved poses from the PR was close to the true pose. This enables successful global localization even if all the preliminarily initialized hypotheses were wrong by periodically initializing new hypotheses based on the retrieved PR pose if computational capacity is available until the null hypothesis probability falls below a threshold. Other ambiguity-aware multi-hypothesis approaches (cf. Sect. II-C) either require GNSS and a vectorized HD map to resolve ambiguities and detect convergence [62], [63], [64], [65], [66] or have to initialize localization hypotheses covering the complete driveable map space [50], [60], [67], [68] leading to a less efficient and scalable approach.

III. CROSS-MODALITY MULTI-HYPOTHESIS LOCALIZATION

Fig. 1 already provided an overview of our proposed GNSS-free cross-modal global multi-hypothesis localization approach. The hypotheses are initialized using a suitable PR method. Subsequently, the goal is to infer which of the initialized hypotheses actually corresponds to the correct one close to the true pose. This is achieved by determining the hypothesis whose odometry measurements and pose observations are most consistent over time. The pose observations are obtained from our deep sensor-to-map matching model which yields as our localization frontend and will be described in Sect. III-A. The intermediate results and representations of this model *for a single hypothesis* are depicted in Fig. 2 for a radar-to-lidar localization experiment on the KAIST02 sequence of the MulRan dataset [3]. Centered around a prior pose hypothesis, a horizontal region is extracted from the map (Fig. 2a) and fed into the model together with the corresponding radar sensor image (Fig. 2b) both represented from a bird's eye view (BEV) perspective. Subsequently, the network brings both the map and the sensor image into common planar embedding spaces (cf. Fig. 2c and Fig. 2d). For three-dimensional data we add a stack of PVConv layers to the input layer which keeps the pipeline for the subsequent modules equivalent, but enables the processing of three-dimensional data. The final layer performs cross-correlations in the ground plane for different discrete yaw orientations whose output corresponds to a probability distribution over possible SE2 poses given the sensor image and the prior pose estimate (Fig. 2e). Note, that these operations are computed for every hypothesis in parallel on the GPU. Next, the correlation output is approximated by a multivariate Gaussian (cf. Sect. III-B) to be further processed by the proposed Gaussian sum filter (GSF) which determines an updated pose estimate for every hypothesis based on the pose observations and the odometry measurements as detailed in Sect. III-C. Additionally, the probability of every hypothesis is updated as well. These steps are then recursively performed whereas a predicted pose of the GSF is used for the map tile cropping in subsequent frames. To account for the case that none of the initialized hypotheses from the PR was correct, we additionally introduce the concept of the null hypothesis [46], [47] (cf. Sect. III-C6) and develop corresponding initialization strategies (cf. Sect. III-C7) that will be employed in symbiosis with PR to be able to find the correct hypothesis during initialization although none of the preliminarily initialized hypotheses was close to the true pose. In the following, we will assume that initial pose estimates have already been provided by a PR method.

TABLE 1. PVCNN preprocessing layer configuration, from bottom (input) layer to top (output) layer.

Block	Layer	Kernel	Channels	Stride
PVConv1	3D Convolution	3	4	1
	MLP	-	4	-
PVConv2	3D Convolution	3	8	1
	MLP	-	8	-
PVConv3	3D Convolution	3	2	1
	MLP	-	2	-

A. LOCALIZATION FRONTEND

The task of the sensor-to-map matching model is to encode localization-relevant features in an embedding space for both sensor and map data. A straightforward way of generating these embeddings is described in the works of [17], which we adapt to our approach. Two encoder-decoder networks with skip connections are used as sensor and map embedding functions. These scan embedding networks are LinkNets [69] using feature maps across four scales with an encoder and decoder block for each scale. Table 2 shows the configuration of the model components in detail Each encoder block is made up of two residual layers, decoder blocks perform upsampling, convolution and addition. The output

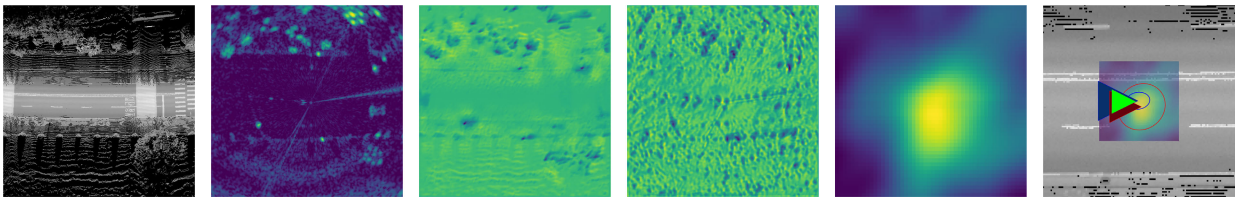


FIGURE 2. Depiction of input, intermediate and output data of the proposed localization approach for one hypothesis for a single frame of the KAIST02 sequence. The depicted correlation image in subfigures (e) and (f) correspond to the most likely yaw angle. In subfigure (f), the *large blue triangle and ellipse* corresponds to the current pose estimate with corresponding 2σ uncertainty ellipse after the Kalman update step, the *medium red triangle and ellipse* corresponds to the pose measurement based on the Gaussian approximation of the correlation surface and the *small green triangle* corresponds to the true pose. The subcaptions comprise the following abbreviations: PC: point cloud, BEV: birds-eye-view, RI: radar image, ME: map embedding, RE: radar embedding, CS: correlation surface.

TABLE 2. Network layer configuration, from bottom (input) layer to top (output) layer.

Block	Layer	Kernel	Channels	Stride
-	Input Conv	7	32	1
Encoder Block 1	Residual	3	32	2
	Residual	3	32	1
Encoder Block 2	Residual	3	32	2
	Residual	3	32	1
Encoder Block 3	Residual	3	64	2
	Residual	3	64	1
Encoder Block 4	Residual	3	128	2
	Residual	3	128	1
Decoder Block 4	Convolution	3	64	1
	Upsample	2	64	2
	Convolution	3	64	1
Decoder Block 3	Convolution	3	32	1
	Upsample	2	32	2
	Convolution	3	32	1
Decoder Block 2	Convolution	3	32	1
	Upsample	2	32	2
	Convolution	3	32	1
Decoder Block 1	Convolution	3	1	1
	Upsample	2	1	2
	Convolution	3	1	1

of the two networks is then cross-correlated in the Fourier domain before fitting a multivariate Gaussian distribution to the output (cf. Sect. III-B). The central change to [17] we perform is an additional stack of three Point Voxel Convolution (PVConv) layers [70] at the input of any network that handles three-dimensional point clouds. This allows the network to distinguish three dimensional objects by assigning similar feature vectors to points that originate from similar elements of the environment. For vehicle localization tasks, commonly only an SE2 pose is estimated under the assumption that the vehicle stays in contact with the ground at all times. Therefore, high resolution along the vertical dimension of the point cloud is not strictly required for the localization task if the features in each horizontal cell are sufficiently distinctive. We project the point cloud onto a two-dimensional grid along the horizontal plane, averaging the features of points that fall into the same cell as shown in Fig. 3. This projection is differentiable with respect to the input features of the points. This layer is in essence a variant of the voxelization applied in [48] with voxels that are infinitely large in the z dimension,

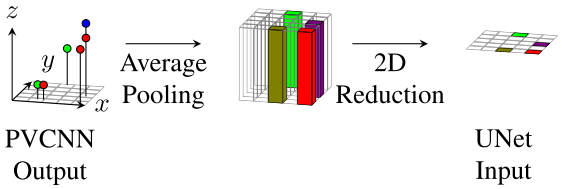


FIGURE 3. Projection procedure for points with a three dimensional feature vector encoded in red, green and blue. Points falling into the same cell will be average pooled, allowing the neural network to perform point-wise weighting.

but with an immediate average pooling of all points within each voxel:

$$\mathbf{f}_{h,w} = \frac{\sum_{\mathbf{p}\in\mathcal{P}} \mathbf{p}_{\text{feature}} \cdot \mathrm{m}(\mathbf{p}, h, w)}{\sum_{\mathbf{p}\in\mathcal{P}} \mathrm{m}(\mathbf{p}, h, w)} \tag{1}$$

with

$$\mathrm{m}(\mathbf{p}, h, w) = \begin{cases} 1, & \text{if } h \le p_y < h + \Delta h \\ & \wedge w \le p_x < w + \Delta w \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{f}_{h,w}$ denotes the feature vector in cell (h, w) of the resulting BEV feature map, $\mathbf{p}_{\text{feature}}$ is the feature vector associated with point $\mathbf{p}$ from point cloud $\mathcal{P}$, p_x and p_y is the x and y position of point $\mathbf{p}$, respectively, and Δh, Δw are vertical and horizontal grid sizes of the BEV feature map. Reducing the dimensionality of the point cloud in this way avoids the memory overhead of a high resolution voxelization in 3D and thereby grants the network the ability to produce a 2D representation of the point cloud at the same resolution that the localization frontend uses for the BEV representation of the sensor data.

This allows the network to learn the importance of 3D features for the SE2 localization task and the characteristics of a particular sensor end-to-end. The network can select the points from the map that a given sensor can detect and vice versa.

The model is trained twice on different datasets to evaluate the performance when localizing across different sensor modalities. The MulRan dataset [3] is used for this purpose as it provides measurements from a radar sensor and a three-dimensional lidar point cloud map. Additionally, we recorded a custom lidar-to-aerial-imagery dataset around

the city of Dortmund, Germany, with our own test vehicle equipped with an Ouster OS1-64 3D lidar sensor and an RTK-INS/GNSS system for GT information. The aerial imagery map is a set of true orthophotos (TrueDOP) provided by the Geobasis NRW [71]. This dataset will be termed Dortmund dataset in the following.

The complete network is trained using Adam [72] with a learning rate of $\eta_{\text{learn}} = 1 \times 10^{-4}$. For regularization of the network, both dropout [73] with a rate of $P_{\text{drop}} = 0.5$ and decoupled weight decay [74] with a rate of $\eta_{\text{decay}} = 1 \times 10^{-6}$ are used. The point clouds are sampled at the input at random with an independently sampled drop probability between Cross entropy loss is used as supervision during training. Learning is stopped once the validation loss stops decreasing with a grace period of 10 epochs. Then, the model with the lowest validation loss is picked for evaluation.

As there might be minor undetected errors in the ground truth of the datasets that might lead to corrupted labels, a technique from learning classification with noisy labels [75] is adapted to the localization task to avoid small errors in the label corrupting the training signal. As computing a full confusion matrix for the search space is prohibitively computationally expensive, a convolution layer with sufficiently small kernel size is introduced instead and initialized with a fixed (i.e. not changing during training) identity weight matrix $\mathbf{D}_{\text{i}}$. A learnable weight matrix $\mathbf{D}_{\text{l}}$ of the same size is initialized with zero weights and added to the fixed identity mapping before the convolution with the correlation output tensor $\mathbf{S}_{\text{corr}}$ is performed:

$$\hat{\mathbf{S}}_{\text{corr}} = \mathbf{S}_{\text{corr}} \star (\mathbf{D}_{\text{l}} + \mathbf{D}_{\text{i}}) \tag{3}$$

Here, $\star$ denotes a 2D convolution. With this configuration, the network can spread out the training signal across a small space in the correlation tensor in the event of a wrong label. $\mathbf{D}_{\text{l}}$ is driven to approximate the unknown inherent error distribution of the poses given in the dataset. Weight decay can be applied to $\mathbf{D}_{\text{l}}$ to discourage the embedding networks from spatially misaligning features. This additional layer is only present during training and is not part of the pose estimation frontend during evaluation.

In our experiments, using this method produced more robust embeddings at the cost of possibly introducing bias into the pose estimate. The pose observations resulting from the network output are therefore calibrated to produce zero bias on the validation set (cf. Sect. IV-C2).

B. GAUSSIAN APPROXIMATION OF THE CORRELATION SURFACE

The GSF assumes that individual measurements, denoted by $\mathbf{z}_k$, follow a zero-mean white-noise process with known covariance $\mathbf{R}_k$. However, the output of our sensor-to-map matching model is a correlation surface over possible SE2 poses. A common approach to obtain a covariance estimate from a correlation surface is the weighted sample covariance method similar to [57]. Here, the mean $\mathbf{z}_k$ of the approximation is set to the maximum of the correlation surface and the covariance is estimated by

$$\mathbf{R}_k = \sum_{\mathbf{s}_k \in \mathbf{S}_{\text{corr},k}} p(\mathbf{s}_k)\,(\mathbf{s}_k - \mathbf{z}_k)\,(\mathbf{s}_k - \mathbf{z}_k)^{\intercal} \tag{4}$$

with $\mathbf{s}_k = [x_{\text{s},k}, y_{\text{s},k}, \psi_{\text{s},k}]^{\intercal}$ being a pose sample and $p(\mathbf{s}_k)$ its relative weight. $\mathbf{S}_{\text{corr},k}$ contains the poses around the predicted pose that are represented by the correlation output tensor. To analyze the goodness of this approximation, we performed an offline per-frame localization on the validation sequence where we added random uniform perturbations to the true poses to obtain a set of prior poses such that the true pose still lies in the search range of the correlation. Based on these perturbed poses the input map tiles are extracted. Together with the corresponding sensor image, we then compute the resulting correlation output tensor and perform the multivariate Gaussian approximation to obtain a pose measurement $(\mathbf{z}_k, \mathbf{R}_k)$ for every perturbed pose. Fig. 4 shows the distribution of the longitudinal errors between the resulting pose measurement means and the true poses. It is visible that the main portion of measurements are approximately Gaussian distributed whereas the rest of the measurements are nearly uniformly distributed along the longitudinal range. Moreover, Fig. 4 reveals that the pose measurements have a non-negligible bias that needs to be accounted for. In the following section, we explicitly consider this error distribution of our frontend network when modeling the measurement likelihood (cf. Sect. III-C3). For this, it is required to explicitly calculate the portion of measurements that approximately follow a Gaussian distribution which can be regarded as the detection probability of our localization frontend. On the other hand, the remaining portion of measurements may be considered as clutter which will be represented in the measurement likelihood as well.

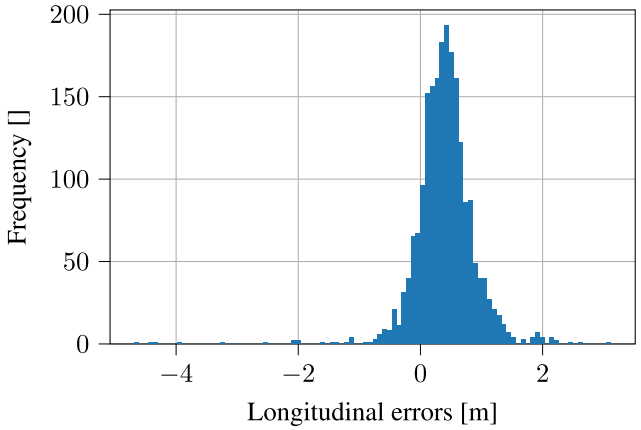


FIGURE 4. Histogram of the longitudinal errors from the offline per-frame localization experiment on the validation sequence.

C. MULTI-HYPOTHESIS EGO-POSE TRACKING

We seek to estimate the location and orientation of the ego-vehicle in UTM coordinates at every time step k, i.e. the 2D pose $\mathbf{x}_k = [x_k \;\; y_k \;\; \psi_k]^{\intercal} \in \mathbb{R}^3$, given a set of observations $Z_{1:k} = \{\mathbf{z}_i\}_{i=1}^{k}$. For online estimation, we use

the recursive Bayesian filtering framework to compute the posterior distribution

$$p\left(\mathbf{x}_k|Z_{1:k}\right) = \eta p\left(Z_k|\mathbf{x}_k\right) p\left(\mathbf{x}_k|Z_{1:k-1}\right) \tag{5}$$

where η is a normalization factor, $p\left(Z_k|\mathbf{x}_k\right)$ the measurement likelihood and

$$p\left(\mathbf{x}_k|Z_{1:k-1}\right) = \int p\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) p\left(\mathbf{x}_{k-1}|Z_{1:k-1}\right) \mathrm{d}\mathbf{x}_{k-1} \tag{6}$$

is the predicted density [5]. Here, $p\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ corresponds to the motion model. Note that if odometry measurements $U_{1:k} = \{\mathbf{u}_i\}_{i=1}^{k}$ with $\mathbf{u}_i = [\delta_{\mathrm{lon},k} \;\; \delta_{\mathrm{lat},k} \;\; \delta\psi_k]^\intercal \in \mathbb{R}^3$ are available, the motion model may be conditioned on the latest odometry measurement $\mathbf{u}_k$, i.e $p\left(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{u}_k\right)$. The main problem when using PR for initialization is that the retrieved pose may be wrong and not in the vicinity of the true pose. Therefore, a self-assessing strategy is required for the localization algorithm to identify if the retrieved pose is actually close to the true pose or completely far off. To tackle this problem, we propose to initially track *multiple* localization hypotheses by initializing the localization algorithm with the top-n poses of the PR. Subsequently, we identify plausible and non-plausible hypotheses over time by determining if the motion and pose observations of an hypothesis are consistent over time based on the corresponding estimated location in the map. Thus, hypotheses where the predicted and observed measurements disagree with each other become less plausible, i.e. their probability decreases, and may be finally pruned if their probability falls below a threshold. To account for the case that none of the initially retrieved poses from the PR is correct, we explicitly estimate this probability based on the concept of a null hypothesis [46], [47] and develop two PR-based initialization strategies (cf. Sect. III-C7).

1) GAUSSIAN MIXTURE REPRESENTATION OF THE BELIEF

There are several possibilities to choose a suitable modeling and representation of the likelihood $p\left(Z_k|\mathbf{x}_k\right)$ and belief $p\left(\mathbf{x}_k|Z_{1:k}\right)$ for a tractable estimation of multiple hypotheses. In this contribution we propose the use of a Gaussian sum filter (GSF) [45], [52] for estimation to keep the number of estimated hypotheses low. This enables us to simultaneously provide our sensor-to-map matching model with a corresponding small batch of map tiles that can be processed in parallel. When employing a GSF for estimation, we assume that the posterior distribution can be represented at every time step k by

$$p\left(\mathbf{x}_k|Z_{1:k}\right) = \sum_{h_k=1}^{n_k} w_k^{(h_k)} \cdot \mathcal{N}\left(\mathbf{x}_k; \boldsymbol{\mu}_{k|k}^{(h_k)}, \boldsymbol{\Sigma}_{k|k}^{(h_k)}\right). \tag{7}$$

Here, $\left\{w_k^{(h_k)}, \boldsymbol{\mu}_{k|k}^{(h_k)}, \boldsymbol{\Sigma}_{k|k}^{(h_k)}\right\}_{h_k=1}^{n_k}$ denote the n_k Gaussian mixture components, termed hypotheses in the following. Each hypothesis is comprised of a weight $w_k^{(h_k)}$, mean $\boldsymbol{\mu}_{k|k}^{(h_k)}$ and covariance $\boldsymbol{\Sigma}_{k|k}^{(h_k)}$ whereas the superscript[1] (h_k) denotes the corresponding index of the hypothesis. The weight of each hypothesis is equal to the estimated probability of the hypothesis representing the true vehicle pose. In the following, we will describe the individual modules of our localization framework. First, we will briefly elaborate on the motion prediction of every hypothesis (cf. Sect. III-C2). Subsequently, based on an evaluation of the pose observations from the multivariate Gaussian approximation of the correlation surface (cf. Sect. III-B), we will explain how we deal with clutter measurements and misdetections using the single-object-tracking formulation of the measurement likelihood [52], [53], [54] (cf. Sect. III-C3) which will be employed to compute the posterior distribution (cf. Sect. III-C4). This includes a Kalman update step for every data association hypothesis as well as the update of the hypotheses weights. In a subsequent step, we will further elaborate on methods that are required for a tractable GSF estimation, namely hypotheses merging, capping and pruning strategies to keep the number of estimated hypotheses low (cf. Sect. III-C5). Finally, we will introduce the concept of the null hypothesis probability (cf. Sect. III-C6) which is accompanied by the proposition of two initialization strategies (cf. Sect. III-C7) to account for the case, that none of the preliminarily initialized hypotheses was correct.

2) PREDICTION

Within the prediction step, the mean of the h_k-th hypothesis is predicted independently using a motion model $\mathbf{f}(\cdot)$ given the prior mean $\boldsymbol{\mu}_{k-1|k-1}^{(h_k)}$ and, if available, the latest motion measurement $\mathbf{u}_k$ according to

$$\boldsymbol{\mu}_{k|k-1}^{(h_k)} = \mathbf{f}\left(\boldsymbol{\mu}_{k-1|k-1}^{(h_k)}, \mathbf{u}_k\right) \tag{8}$$

If motion measurements $\mathbf{u}_k$ are available, the employed motion model depends on the type of measurements quantifying the motion [5]. For odometry measurements, i.e. $\mathbf{u}_k = \left[\delta_{\mathrm{lon},k} \;\; \delta_{\mathrm{lat},k} \;\; \delta_{\psi,k}\right]^\intercal$, the predicted mean is computed by [76]

$$\begin{bmatrix} x_{k|k-1}^{(h_k)} \\ y_{k|k-1}^{(h_k)} \\ \psi_{k|k-1}^{(h_k)} \end{bmatrix} = \begin{bmatrix} x_{k-1|k-1}^{(h_k)} \\ y_{k-1|k-1}^{(h_k)} \\ \psi_{k-1|k-1}^{(h_k)} \end{bmatrix} + \begin{bmatrix} \delta_{\mathrm{lon},k} \cos\left(\psi_{k-1|k-1}^{(h_k)}\right) - \delta_{\mathrm{lat},k} \sin\left(\psi_{k-1|k-1}^{(h_k)}\right) \\ \delta_{\mathrm{lon},k} \sin\left(\psi_{k-1|k-1}^{(h_k)}\right) + \delta_{\mathrm{lat},k} \cos\left(\psi_{k-1|k-1}^{(h_k)}\right) \\ \delta_{\psi,k} \end{bmatrix} \tag{9}$$

[1] Note, that this index notation will be used throughout in the following, i.e. a superscript inside brackets $(\cdot)$ denotes an index and not an exponent. Vice versa, a superscript without brackets denotes an exponent.

Here, $\delta_{\text{lon},k}$ and $\delta_{\text{lat},k}$ correspond to the measured longitudinal and lateral translation in the vehicle-fixed frame from time step $k-1$ to time step k whereas $\delta_{\psi,k}$ denotes the measured relative rotation in the yaw angle.

If the motion measurements are provided in the form of translational and angular velocity, i.e. $\mathbf{u}_k = [v_k \ \omega_k]^\intercal$, the constant turn rate velocity (CTRV) model [77] may be employed to determine the predicted mean according to

$$\begin{bmatrix} x_{k|k-1}^{(h_k)} \\ y_{k|k-1}^{(h_k)} \\ \psi_{k|k-1}^{(h_k)} \end{bmatrix} = \begin{bmatrix} x_{k-1|k-1}^{(h_k)} \\ y_{k-1|k-1}^{(h_k)} \\ \psi_{k-1|k-1}^{(h_k)} \end{bmatrix} + \begin{bmatrix} \frac{v_k}{\omega_k}\left(-\sin\left(\psi_{k-1|k-1}^{(h_k)}\right)+\sin\left(\psi_{k-1|k-1}^{(h_k)}+\omega_k\Delta t\right)\right) \\ \frac{v_k}{\omega_k}\left(\cos\left(\psi_{k-1|k-1}^{(h_k)}\right)-\cos\left(\psi_{k-1|k-1}^{(h_k)}+\omega_k\Delta t\right)\right) \\ \omega_k\Delta t \end{bmatrix} \tag{10}$$

Here, Δt denotes the time between frames $k-1$ and k in seconds. The predicted covariance $\mathbf{\Sigma}_{k|k-1}^{(h_k)}$ is given by

$$\mathbf{\Sigma}_{k|k-1}^{(h_k)} = \mathbf{F}_k \mathbf{\Sigma}_{k-1|k-1}^{(h_k)} \mathbf{F}_k^\intercal + \mathbf{V}_k \mathbf{M}_k \mathbf{V}_k^\intercal \tag{11}$$

where

$$\mathbf{F}_k = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\boldsymbol{\mu}_{k-1|k-1}^{(h_k)}, \mathbf{u}_k}, \quad \mathbf{V}_k = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right|_{\boldsymbol{\mu}_{k-1|k-1}^{(h_k)}, \mathbf{u}_k} \tag{12}$$

denote the state and motion noise Jacobian matrices respectively. The matrix $\mathbf{M}_k$ denotes the motion noise covariance matrix in the corresponding motion measurement space and is usually defined by a diagonal matrix [5]. This yields $\mathbf{M}_k = \text{diag}\left(\sigma_{\delta_{\text{lon}}}^2, \sigma_{\delta_{\text{lat}}}^2, \sigma_{\delta_\psi}^2\right)$ for the odometry and $\mathbf{M}_k = \text{diag}\left(\sigma_v^2, \sigma_\omega^2\right)$ for the CTRV motion model.

3) MEASUREMENT LIKELIHOOD

The evaluation of the Gaussian approximation of the output of our deep sensor-to-map matching model revealed (cf. Sect. III-B) that only a portion of the resulting measurements are actually Gaussian-distributed around the true pose (assuming that an occurring bias has been corrected) while the rest of the measurements correspond to clutter which is approximately uniformly distributed within the search range. Therefore, if a measurement is received, the data association is unknown, i.e. it is primarily obscure if the measurement corresponds to the true pose or clutter. To model misdetections and clutter among unknown data association, we employ the single-object tracking (SOT) formulation of the complete measurement likelihood [52], [53], [54]

$$p(Z_k|\mathbf{x}_k) = \left[(1-p_\text{d}) + \frac{p_\text{d}}{\lambda_\text{c}} \sum_{\theta_k=1}^{m_k} \mathcal{N}\left(\mathbf{z}_k^{(\theta_k)}; \mathbf{x}_k, \mathbf{R}_k^{(\theta_k)}\right)\right] \cdot \frac{e^{-\lambda_\text{c} V}}{m_k!} (\lambda_\text{c} V)^{m_k}. \tag{13}$$

Here, $\theta_k \in \mathbb{N}_0$ denotes the data association variable. If $\theta_k \in \{1, 2, \ldots, m_k\}$ it implies that $\mathbf{z}_k^{(\theta_k)}$ is a measurement of the true vehicle pose, i.e. the ego-vehicle pose has been detected. In this case we assume that the spatial likelihood function is Gaussian with covariance $\mathbf{R}_k^{(\theta_k)}$ (cf. Sect. III-B). Otherwise, $\theta_k = 0$ corresponds to a misdetection. Here, p_d denotes the probability of detecting the ego-pose. Based on the evaluation of our Gaussian approximation of the correlation output, we model the detection probability as a constant and estimate its value by the portion of measurements that actually follow a Gaussian distribution around the true pose. On the other hand, the clutter measurements are assumed to follow a uniform distribution over the correlation search volume $V = a_\text{search}^2 \psi_\text{search}$. Here, a_search corresponds to the search range in both translational directions and ψ_search to the search range in yaw direction. The number of clutter measurements is assumed to be Poisson distributed with spatial density λ_c. The spatial density may be defined as $\lambda_\text{c} = \bar{\lambda}_\text{c}/V$, where $\bar{\lambda}_\text{c}$ corresponds to the expected number of clutter measurements per frame. An estimation of both the detection probability p_d and the clutter rate $\bar{\lambda}_\text{c}$ will be described in Sect. IV-C2.

4) COMPUTATION OF THE POSTERIOR DISTRIBUTION

For the computation of the posterior distribution $p(\mathbf{x}_k|Z_{1:k})$, we can now plug (13) back into (5) and incorporate the term outside the bracket of (13) in the normalizer η. Since the predicted density $p(\mathbf{x}_k|Z_{1:k-1})$ is a Gaussian sum, the computation of $p(\mathbf{x}_k|Z_{1:k})$ only involves the multiplication of Gaussians which yields another Gaussian sum with $\tilde{n}_k = n_{k-1}(m_k+1)$ components [53]. The index $\tilde{h}_k$ of the $\tilde{h}_k$-th posterior component is given by

$$\tilde{h}_k = h_{k-1} + n_{k-1}\theta_k \tag{14}$$

where $h_{k-1} \in \{1, 2, 3, \ldots, n_{k-1}\}$ denotes the index of the prior hypothesis. The index set $\tilde{I}_k = \{1, 2, 3, \ldots, \tilde{n}_k\}$ comprises the indices of all posterior hypotheses, i.e. $\tilde{h}_k \in \tilde{I}_k$. The computation of the posterior mean $\boldsymbol{\mu}_{k|k}^{(\tilde{h}_k)}$ and covariance $\mathbf{\Sigma}_{k|k}^{(\tilde{h}_k)}$ differs based on the corresponding value of the data association variable θ_k. If $\theta_k > 0$, measurement $\mathbf{z}_k^{(\theta_k)}$ is associated to the h_{k-1}-th hypothesis and the posterior mean and covariance are computed by the EKF update equations

$$\mathbf{S}_k^{(\tilde{h}_k)} = \mathbf{H}_k^{(h_{k-1})} \mathbf{\Sigma}_{k|k-1}^{(h_{k-1})} \left(\mathbf{H}_k^{(h_{k-1})}\right)^\intercal + \mathbf{R}_k^{(\theta_k)} \tag{15}$$

$$\mathbf{K}_k^{(\tilde{h}_k)} = \mathbf{\Sigma}_{k|k-1}^{(h_{k-1})} \left(\mathbf{H}_k^{(h_{k-1})}\right)^\intercal \left(\mathbf{S}_k^{(\tilde{h}_k)}\right)^{-1} \tag{16}$$

$$\boldsymbol{\nu}_k^{(\tilde{h}_k)} = \mathbf{z}_k^{(\theta_k)} - \mathbf{h}\left(\boldsymbol{\mu}_{k|k-1}^{(h_{k-1})}\right) \tag{17}$$

$$\boldsymbol{\mu}_{k|k}^{(\tilde{h}_k)} = \boldsymbol{\mu}_{k|k-1}^{(h_{k-1})} + \mathbf{K}_k^{(\tilde{h}_k)} \boldsymbol{\nu}_k^{(\tilde{h}_k)} \tag{18}$$

$$\mathbf{\Sigma}_{k|k}^{(\tilde{h}_k)} = \left(\mathbf{I} - \mathbf{K}_k^{(\tilde{h}_k)} \mathbf{H}_k^{(h_{k-1})}\right) \mathbf{\Sigma}_{k|k-1}^{(h_{k-1})} \tag{19}$$

where the measurement model is given by

$$\mathbf{h}(\mathbf{x}) = \begin{bmatrix} x + b_{\text{lon}}\cos(\psi) - b_{\text{lat}}\sin(\psi) \\ y + b_{\text{lon}}\sin(\psi) + b_{\text{lat}}\cos(\psi) \\ \psi + b_{\psi} \end{bmatrix}. \tag{20}$$

Here, b_{lon}, b_{lat} and b_{ψ} denote optional bias terms in the longitudinal and lateral direction as well as in the yaw orientation that may be incorporated into the measurement model if observed (cf. Sect.III-B). The measurement Jacobian $\mathbf{H}$ then results in

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & -b_{\text{lon}}\sin(\psi) - b_{\text{lat}}\cos(\psi) \\ 0 & 1 & b_{\text{lon}}\cos(\psi) - b_{\text{lat}}\sin(\psi) \\ 0 & 0 & 1 \end{bmatrix} \tag{21}$$

whereas $\mathbf{H}_k^{(h_{k-1})}$ means that the Jacobian is evaluated at $\boldsymbol{\mu}_{k|k-1}^{(h_{k-1})}$. For $\theta_k = 0$, i.e. a misdetection, no Kalman update is performed and the posterior mean and covariance are equal to the predicted mean $\boldsymbol{\mu}_{k|k-1}^{(h_{k-1})}$ and covariance $\boldsymbol{\Sigma}_{k|k-1}^{(h_{k-1})}$ respectively. Finally, the *unnormalized* weights $\tilde{w}_{k|k}^{(\tilde{h}_k)}$ of the posterior mixture components are calculated by [45], [52], and [53]

$$\tilde{w}_k^{(\tilde{h}_k)} = \begin{cases} w_k^{h_{k-1}}(1 - p_{\text{d}}) & \text{if } \theta_k = 0 \\ w_k^{h_{k-1}} \dfrac{p_{\text{d}}}{\lambda_{\text{c}}} \mathcal{N}\left(\mathbf{z}_k^{(\theta_k)}; \mathbf{h}\left(\boldsymbol{\mu}_{k|k-1}^{(h_{k-1})}\right), \mathbf{S}_k^{(\tilde{h}_k)}\right) & \text{if } \theta_k > 0 \end{cases} \tag{22}$$

Finally, to obtain a valid posterior pdf, we normalize $\tilde{w}_k^{(\tilde{h}_k)}$ over all n_k components

$$w_k^{(\tilde{h}_k)} = \frac{\tilde{w}_k^{(\tilde{h}_k)}}{\sum_{i=1}^{\tilde{n}_k} \tilde{w}_k^{(i)}}. \tag{23}$$

5) HYPOTHESES MERGING, CAPPING AND PRUNING

The formulation of the GSF is, without any further post-processing, intractable as the number of estimated hypotheses would grow exponentially over time with $\prod_{i=1}^{k}(m_i + 1)$. Therefore, there is a need for limiting the number of hypotheses. Common concepts for this purpose are merging, capping and pruning of hypotheses. For this, we follow a combination of the approaches of [46], [47], and [78]: first, we determine the most likely hypothesis of the current set of unprocessed hypotheses. If this hypothesis resulted from an association with a measurement, i.e. its index $j > n_{k-1}$ (cf. (14)), we merge it with all nearby posterior hypotheses which resulted likewise from an association to a measurement, i.e. $i > n_{k-1}$, using the same moment-matching procedure as proposed in [78]. Here, nearby means that the Mahalanobis distance between the hypotheses is below a threshold τ_{merge}. Additionally, if the posterior hypothesis which resulted from a misdetection ($\theta_k = 0$) of the same prior hypothesis as the j-th hypothesis, i.e. $i_{\theta_k=0} = (j \mod n_{k-1})$ (cf. (14)), is nearby as well, we add its weight to the weight of the merged posterior hypotheses. If the j-th hypothesis resulted from a misdetection, it is kept and not merged with any other hypotheses. This extends the merging approach of [78] by the "hypotheses splitting" concept proposed by [47] with the difference that the splitted hypotheses do not have equal weights, but the weight ratio is determined by the corresponding detection probability p_{d} and clutter intensity λ_{c} respectively. Details of our merging concept are depicted in Algorithm 1.

Algorithm 1 Hypotheses Merging Algorithm

Require: $\tilde{I}_k, \left\{w_k^{(\tilde{h}_k)}, \boldsymbol{\mu}_{k|k}^{(\tilde{h}_k)}, \boldsymbol{\Sigma}_{k|k}^{(\tilde{h}_k)}\right\}_{\tilde{h}_k=1}^{\tilde{n}_k}$

$l = 0$
repeat
  $l = l + 1$
  $j = \arg\max_{i \in \tilde{I}_k} w_k^{(i)}$
  if $j > n_{k-1}$ **then**
    $M = \left\{ i \in \tilde{I}_k \,\middle|\, \left\| \boldsymbol{\mu}_{k|k}^{(i)} - \boldsymbol{\mu}_{k|k}^{(j)} \right\|_{\boldsymbol{\Sigma}_{k|k}^{(i)}} \leq \tau_{\text{merge}} \wedge i > n_{k-1} \right\}$
    $\bar{w}_{k|k}^{(l)} = \sum_{i \in M} w_{k|k}^{(i)}$
    $\bar{\boldsymbol{\mu}}_{k|k}^{(l)} = \frac{1}{\bar{w}_{k|k}^{(l)}} \sum_{i \in M} w_{k|k}^{(i)} \boldsymbol{\mu}_{k|k}^{(i)}$
    $\bar{\boldsymbol{\Sigma}}_{k|k}^{(l)} = \frac{1}{\bar{w}_{k|k}^{(l)}} \sum_{i \in M} w_{k|k}^{(i)} \left[\boldsymbol{\Sigma}_{k|k}^{(i)} + \left(\bar{\boldsymbol{\mu}}_{k|k}^{(l)} - \boldsymbol{\mu}_{k|k}^{(i)}\right)\left(\bar{\boldsymbol{\mu}}_{k|k}^{(l)} - \boldsymbol{\mu}_{k|k}^{(i)}\right)^{\mathsf{T}} \right]$
    $i_{\theta_k=0} = (j \mod n_{k-1})$
    if $\left\| \boldsymbol{\mu}_{k|k}^{(i_{\theta_k=0})} - \boldsymbol{\mu}_{k|k}^{(j)} \right\|_{\boldsymbol{\Sigma}_{k|k}^{(i_{\theta_k=0})}} \leq \tau_{\text{merge}}$ **then**
      $\bar{w}_{k|k}^{(l)} = \bar{w}_{k|k}^{(l)} + w_k^{(i_{\theta_k=0})}$
    end if
  else
    $M = \{j\}$
    $\bar{w}_{k|k}^{(l)} = w_k^{(j)}, \bar{\boldsymbol{\mu}}_{k|k}^{(l)} = \boldsymbol{\mu}_{k|k}^{(j)}, \bar{\boldsymbol{\Sigma}}_{k|k}^{(l)} = \boldsymbol{\Sigma}_{k|k}^{(j)}$
  end if
  $\tilde{I}_k = \tilde{I}_k \setminus M$
until $\tilde{I}_k = \emptyset$

Subsequently, we cap the $\bar{n}_k$ merged posterior hypotheses $\left\{\bar{w}_k^{(\bar{h}_k)}, \bar{\boldsymbol{\mu}}_{k|k}^{(\bar{h}_k)}, \bar{\boldsymbol{\Sigma}}_{k|k}^{(\bar{h}_k)}\right\}_{\bar{h}_k=1}^{\bar{n}_k}$ to $n_{\max}$ most likely hypotheses and normalize the hypotheses weights. Finally, we prune all hypotheses with weights $\bar{w}_k^{(\bar{h}_k)} < \tau_{\text{prune}}$ and renormalize if necessary to obtain the posterior mixture components $\left\{w_k^{(h_k)}, \boldsymbol{\mu}_{k|k}^{(h_k)}, \boldsymbol{\Sigma}_{k|k}^{(h_k)}\right\}_{h_k=1}^{n_k}$.

So far we described how to predict and update the weight, mean and covariance of multiple localization hypotheses given pose observations from our localization frontend model and available odometry measurements. We further elaborated on the problem that the number of localization hypotheses would grow exponentially over time when we explicitly consider that our frontend may misdetect the ego-pose and produce clutter measurements. To keep the number of

localization hypotheses limited and thus the GSF framework tractable, we introduced merging, capping and pruning strategies in this section. The following two sections elaborate on one remaining task that still needs to be discussed to obtain a complete localization system, namely the initialization of the localization hypotheses, in our case, based on a given PR method.

6) INITIALIZATION WITH INTRODUCTION OF THE NULL HYPOTHESIS

As previously discussed we propose a PR-based initialization in this contribution to obtain a complete GNSS-free vehicle localization system. Since our approach is capable of processing and estimating multiple localization hypotheses we can initialize our system with the top-n retrieved poses from the PR. However, there remains the problem that we cannot be sure if one of the initialized hypotheses will actually lie near the true pose since the recall of a PR method is usually below 100 % even for larger n [1], [35]. Hence, when initializing the GSF with the top-n retrieved poses, there is a non-negligible chance that none of the retrieved will be close to the true pose. To tackle this problem, we explicitly incorporate and estimate this probability within our GSF framework by introducing a so-called *null hypothesis*. This approach is inspired by Jensfelt et al. [46], [47] who introduced the concept of a null hypothesis within a multi-hypothesis global localization approach for an indoor mobile robot. The null hypothesis can be viewed as a uniform distribution assigning a probability of $\hat{w}_k^{(0)}$ to all possible vehicle states. Therefore, before initialization, i.e. at $k = 0$,

$$\hat{w}_0^{(0)} = 1 \tag{24}$$

is satisfied. Now, if we initialize hypotheses, the null hypothesis probability decreases by the probability that the true pose is found among the initialized hypotheses [47]. We denote this probability by $\Pr_n(F)$ where F corresponds to the event that the true pose is found among the n initialized hypotheses and $\neg F$ to its complement. The calculation of $\Pr_n(F)$ is twofold. On the one hand, it depends on the top-n recall $\mathrm{r_{PR}}(n)$ of the PR method which is associated with the probability that *at least one* of the $n \in \mathbb{N}$ retrieved poses will be near the true pose. Since we require not only initial pose estimates, but also corresponding uncertainties for initializing the GSF, we pass the retrieved poses from the PR preliminarily through our localization frontend by extracting n map tiles centered around the retrieved poses and provide them as a batch input to the model. However, our localization frontend detects the true pose only with a probability of p_d which has to be incorporated in the null hypothesis calculation as well. Let B_k denote the event that *exactly* k out of the top-n retrieved PR poses actually lie in range of the true pose. Furthermore, let $\Pr_n(\neg F|B_k)$ denote the conditional probability that the true pose is *not found* given that k out of n the retrieved poses lie in range of the true pose. Now, assuming that the detection of the true pose given different retrieved PR poses is independent of each other, we can calculate $P_n(\neg F|B_k)$ by

$$P_n(\neg F|B_k) = (1 - p_\mathrm{d})^k. \tag{25}$$

Since the events B_k for $k = 0, 1, 2, \ldots, n$ are pairwise disjoint and $\sum_{k=0}^{n} B_k = 1$, we can employ the law of total probability to compute $P_n(\neg F)$ such that

$$\begin{aligned} P_n(\neg F) &= \sum_{k=0}^{n} P_n(\neg F|B_k) P_n(B_k) \\ &= \sum_{k=0}^{n} (1 - p_\mathrm{d})^k P_n(B_k). \end{aligned} \tag{26}$$

Therefore, after initializing n new hypotheses based on the top-n retrieved PR poses, we decrease the null hypothesis probability by

$$\begin{aligned} \Delta\hat{w}_k^{(0)} &= \hat{w}_k^{(0)} \left(1 - P_n(\neg F)\right) \\ \hat{w}_{k+1}^{(0)} &= \hat{w}_k^{(0)} - \Delta\hat{w}_k^{(0)} \end{aligned} \tag{27}$$

whereas the probability mass $\Delta\hat{w}_k^{(0)}$ is distributed across the weights $\hat{w}_k^{(h_k)}$ of the initialized hypotheses such that

$$\hat{w}_k^{(0)} + \sum_{h_k=1}^{n_k} \hat{w}_k^{(h_k)} = 1 \tag{28}$$

is satisfied at every time step k (cf. [46], [47]). The approach therefore requires knowledge about $P_i(B_k)$ for $i = 0, 1, 2, \ldots, n$ and $k = 0, 1, 2, \ldots, i$ whereas $P_n(B_0)$ corresponds to $1 - \mathrm{r_{PR}}(n)$ of the employed PR method. We estimate $P_i(B_k)$ based on the evaluation of the PR model on the validation set and empirically show the suitability of this approach through experiments.

Additionally, to account for the fact that the retrieved poses from the PR are ranked based on a similarity score, it is possible to further weight them relatively to this score. In our case the PR method returns an L_2 distance metric $d_k^{(h_k)}$ for every retrieved pose such that we can weight the initialized hypotheses by their inverse distances according to

$$\hat{w}_k^{(h_k)} = \Delta\hat{w}_k^{(0)} \cdot \frac{\left(1/d_k^{(h_k)}\right)}{\sum_{i=1}^{n_k} \left(1/d_k^{(i)}\right)} \tag{29}$$

Note, that all previous calculations involving the hypotheses weights $w_k^{(h_k)}$ remain valid and those weights are related through the null hypothesis probability $\hat{w}_k^{(0)}$ by

$$w_k^{(h_k)} = \frac{\hat{w}_k^{(h_k)}}{1 - \hat{w}_k^{(0)}}, h_k \in \{1, 2, 3, \ldots, n_k\}. \tag{30}$$

Finally, whenever a hypothesis is pruned (cf. Sect III-C5), its weight is added to $\hat{w}_k^{(0)}$ as proposed by [46] and [47].

7) INITIALIZATION STRATEGIES

Through the introduction of the null hypothesis, it is now possible to explicitly estimate the probability that none of the tracked hypotheses by the GSF is actually close to the true pose. With (27), we further have a formulation for decreasing the null hypothesis probability when initializing new hypotheses. Therefore, a general initialization strategy would be to initialize new hypotheses until the null hypothesis probability $\hat{w}_k^{(0)}$ falls below an acceptable predefined threshold τ_{init}. However, the problem remains that there is an upper bound $n_{\max}$ for the number of hypotheses that can be simultaneously processed by our localization frontend due to computational limits. The goal therefore is to keep the number of estimated hypotheses upper bounded while simultaneously decreasing the null hypothesis probability. To achieve this goal, we can further profit from the GSF framework for localization since the number estimated hypotheses may decrease over time through hypotheses merging or pruning (cf. Sect. III-C5). We therefore propose to continuously retrigger the PR over time to initialize new possible hypotheses whenever computational resources are available, i.e. $n_k < n_{\max}$, until $\hat{w}_k^{(0)} < \tau_{\text{init}}$. With this concept in mind, we propose a *greedy* as well as a *conservative* initialization strategy which will be described in the following:

- *Greedy*: this strategy is a straightforward implementation of the previously described concept, i.e. whenever $n_k < n_{\max}$, the PR is retriggered and $n_{\text{init}} = n_{\max} - n_k$ new hypotheses are initialized.
- *Conservative*: this strategy is motivated by the reasoning that as long as there is more than one hypothesis probable, i.e. $n_k > 1$, there is no need to initialize new hypotheses as the localization system should be unavailable anyways since the current scenario is ambiguous and the most likely hypothesis might be wrong. This reasoning is inspired by [66] who recommend to not ''use'' the most likely localization estimate if there exist multiple probable hypotheses. This concept can be seen as an additional integrity check. Therefore, when using the conservative initialization strategy, the PR is only retriggered if $n_k = 1$ such that always $n_{\text{init}} = n_{\max} - 1$ new hypotheses are initialized.

8) LOCALIZATION ESTIMATE AND AVAILABILITY

Given all acquired information from sensor and map data, the purpose of a localization system is to eventually output an estimate of the ego-vehicle pose in the map. The most popular estimator in the field of localization is the maximum-a-posteriori (MAP) estimate, i.e. the state $\mathbf{x}_k$ which maximizes $p(\mathbf{x}_k|Z_{1:k})$ (cf. (5)) [79]. For a GSF, a reasonable approximation of the MAP estimate is the the mean of the hypothesis with maximum weight [52], i.e.

$$\mathbf{x}_k^{\text{MAP}} = \boldsymbol{\mu}_{k|k}^{(j)}, \quad j = \underset{i=1,2,3,\ldots,n_k}{\arg\max}\ w_k^{(i)}. \tag{31}$$

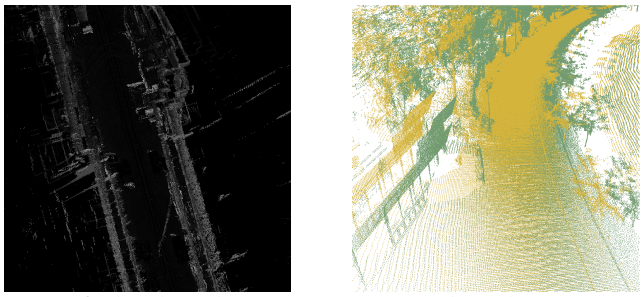
However, localization systems for automated vehicles are required to provide means of self-assessment and integrity checks [43], [63], [66], [80] to avoid dispensing misleading information. In terms of localization, misleading information means that the localization system is confident about its currently estimated location whereas the actual localization error is larger than estimated and even exceeds safety-critical margins [44]. One error source for a localization system to dispense misleading information is ambiguity [63], [66]. Since one purpose of a multi-hypothesis localization approach is to discover ambiguous situations, special caution has to be taken if more than one hypothesis is probable. Inspired by [47] and [66], we therefore propose the following two conditions that need to be satisfied for our localization system to be available. On the one hand, we propose that the null hypothesis probability has to be below a predefined threshold, i.e. $\hat{w}_k^{(0)} < \tau_{\text{init}}$. Note, that τ_{init} is a tuning parameter which trades off the convergence time with robustness. The second condition is that only one estimated hypothesis remains, i.e. $n_k = 1$, such that the probability of an ambiguous scenario is low. In all other cases, the localization system is unavailable.

IV. EXPERIMENTS

To evaluate our approach, we perform experiments for two datasets that differ both in their sensor and map modalities. Our first experiments are performed on the MulRan dataset [3]. Here, the task is to localize a two-dimensional radar sensor image in a three-dimensional lidar point cloud map. We explicitly evaluate the capability of our approach to globally localize the vehicle without any GNSS prior by inferring the correct hypothesis over time from sets of hypotheses that have been initialized by a PR algorithm. For this, we employ a state-of-the-art radar-to-lidar PR algorithm by [1]. We compare our results with the state-of-the-art metric radar-to-lidar localization method RaLL [2]. Here, we also utilize the same odometry estimates provided by the radar odometry from the RaLL method. To show that our approach is likewise suited for different sensor-to-map modalities, we evaluate our approach on a custom lidar-to-aerial-imagery dataset around the city of Dortmund, Germany. The evaluated routes comprise highway and urban scenarios. In the following, we discuss why we opted for the MulRan dataset for radar-to-lidar localization opposed to other popular datasets like the Oxford Radar Robotcar dataset [81] for example.

A. PROBLEMS WITH CURRENT RADAR-TO-LIDAR DATASETS FOR MAP-BASED LOCALIZATION

The authors of RaLL [2] trained their model on the Oxford Radar Robotcar dataset [81] and evaluated their localization accuracy on the same dataset as well as on the MulRan dataset [3]. However, during the training and evaluation of our model we realized that there are several problems / limitations with both datasets for the evaluation of map-based localization algorithms. One requirement for supervised training and testing is the availability of a globally consistent map or set of maps as well as GT poses both accurately referenced to a fixed world coordinate system. When regarding Oxford sequences, we realized that the GT poses as well as the maps



(a) Oxford Radar Robotcar (b) MulRan

FIGURE 5. Visualization of the inconsistencies in the ground truth poses and lidar point cloud maps in both the MulRan and the Oxford Radar Robotcar dataset. Subfigure (a) shows the inconsistencies in the map from a single drive. Here, buildings at the road side are duplicated possibly as a result of inaccurate loop closure. Subfigure (b) shows the inconsistencies between the KAIST01 (*yellow*) and KAIST02 (*green*) point cloud map where the offsets are visible in the building facade on the left side and the tree trunks on the right side.

show global inconsistencies probably resulting from inaccurate loop closures as visible in Fig. 5a. These inconsistencies lead to wrong labels for the training of the localization model thus yielding a degraded performance. Odometry estimation, for which the Oxford dataset has been predominantly used so far (cf. [58], [82], [83], [84], [85], [86]), only requires relative GT and map consistency which is satisfied by the dataset. In the case of MulRan, the maps and GT poses of the individual sequences are in itself consistent, however they are not globally consistent anymore when regarding the maps and GT poses from two traversals of the same scenario but different sequences (cf. Fig. 5b). Therefore, creating the map based on a drive of one sequence and performing localization with a drive from another sequence of the same scenario, as proposed by [2], is problematic as the corresponding GT poses are not valid anymore. Hence, to reasonably compare our approach with RaLL we only evaluate the radar-to-lidar localization results on the KAIST02, DCC01 and Riverside02 sequences from MulRan since here both the mapping and localization drive from RaLL originated from the same sequence such that the localization results are meaningful. For training and validation, we use the Sejong01 drive of the MulRan dataset since it is in itself consistent and was recorded in a completely different environment with respect to the test sequences.

B. INPUT DATA

For the MulRan experiments, the input data is comprised of 360° radar sensor images, the radar odometry estimates of the RaLL method [2] as well as the lidar point cloud map for the corresponding sequences. The localization frame rate is determined by the measurement rate of the radar sensor and corresponds to $\Delta t = 0.25\,\mathrm{s}$.

Regarding the Dortmund dataset, the lidar sensor data is recorded with an Ouster OS1-64 with 10 Hz which equally corresponds to the frame rate of the localization algorithm in this case. Furthermore, the yaw rate measurements are provided by the lidar's internal IMU. To reduce noise, a moving average filter has been applied to the yaw rate measurements. The velocity measurements are received from the CAN bus from the Nissan Leaf ZE0 test vehicle. Ground truth poses are obtained from a RTK-GNSS/INS reference system. Before passing the lidar point cloud to the localization algorithm, it is motion compensated using the measured odometry. The aerial imagery map is a set of true orthophotos (TrueDOP) provided by the Geobasis NRW [71] and is publicly available.[2] True orthophotos eliminate the parallax effects of a camera mounted on an airplane to project pixels onto geographic coordinates [87] by fusing multiple images taken of the same area, providing a consistent map of all ground points that are not vertically occluded. The orthophotos used in this work have a ground sampling distance of 10 cm px^{-1}. The mean absolute positioning error of the orthophotos is stated by [71] as 20 cm to 30 cm. The model is trained on a highway segment of the A45 as well as the drive to and from the highway through low density urban areas with a total length of 24 km. As test scenarios we chose a highway segment of the A40 of 18.73 km length as well as an urban route of 2.92 km length, both close to TU Dortmund University.

C. PARAMETER CALIBRATION

One problem that persists through many localization approaches is that of calibrating the uncertainty estimation. This amounts to estimating covariance matrices that are *consistent*, i.e. accurately capture the expected error that the system produces in the real world. We therefore aim to estimate the necessary calibration parameters systematically from data. First, the estimates of the scan-to-map-matching models are calibrated, as deep neural networks tend to be overconfident in their estimates [88]. Based on the results of this frontend calibration, the parameters for the measurement likelihood can be empirically derived from data as well. To avoid influence on the evaluation, calibration is performed on the validation data set for both the MulRan and Dortmund data.

1) CALIBRATING THE LOCALIZATION FRONTEND

The calibration of the network output results in the calibration of the temperature parameter β of the softmax function. For this, we perform a per-frame localization experiment on the validation set. The network is configured to use a translational search space of 5 m by 5 m with $\pm 15°$ in yaw orientation, i.e. $a_{\mathrm{search}} = 5\,\mathrm{m}$ and $\psi_{\mathrm{search}} = 30°$ (cf. Sect. III-C3), representing the maximum search space that will be evaluated in this work. The translational cell size corresponds to 0.1 m px^{-1} whereas the yaw range is divided into 21 cells with a suitable spacing that will be specified later for the employed method.

In order to generate realistic measurement covariances for the EM, the network is calibrated as in [58] by evaluating an offline validation scenario. Here, the initial guess for the pose is placed according to $\mathbf{x}_{\mathrm{init}} = \mathbf{x}_{\mathrm{GT}} \oplus \mathbf{x}_{\mathrm{random}}$ where $\mathbf{x}_{\mathrm{GT}}$ denotes the ground truth (GT) pose and $\mathbf{x}_{\mathrm{random}}$ is

[2]https://www.geoportal.nrw

independently sampled from a zero-mean multivariate Gaussian distribution with covariance diag $\left((\frac{2.5\,\mathrm{m}}{3})^2\ (\frac{2.5\,\mathrm{m}}{3})^2\ (\frac{5^\circ}{3})^2\right)$ for each example of one test drive that was part of the validation set. $\oplus$ is the pose composition operator [89]. This offsets the initial guess such that approximately 99% of samples will be within the search space. As opposed to the network training, the normal distribution is selected for sampling the map misalignment here to mimic what is to be expected from the online localization method when using the previous time step's estimate as initial guess. The softmax function

$$\mathrm{softmax}_i(\mathbf{S}_{\mathrm{corr}}, \beta) = \frac{\exp(\frac{s_i}{\beta})}{\sum_{k=1}^{|\mathbf{S}_{\mathrm{corr}}|} \exp(\frac{s_k}{\beta})} \tag{32}$$

converts bin s_i of the correlation output $\mathbf{S}_{\mathrm{corr}}$. In the general case, $\mathbf{S}_{\mathrm{corr}}$ does not represent a probability density to the i-th bin of a discretized approximation of the underlying PDF. The temperature parameter β is therefore selected to scale the mean of the squared Mahalanobis distance

$$d_k^2 = (\hat{\boldsymbol{\mu}}_k - \mathbf{x}_{\mathrm{GT},k})^\intercal \hat{\mathbf{R}}_k^{-1} (\hat{\boldsymbol{\mu}}_k - \mathbf{x}_{\mathrm{GT},k}) \tag{33}$$

from the distribution maximum to the ground truth, i.e. $n_{\mathrm{dim}} = \overline{d^2} = \frac{1}{K}\sum_{k=1}^{K} d_k^2$ [58]. $\hat{\mathbf{R}}_k$ is the covariance computed from weighting all cells in the correlation output as defined in (4). The selection of β for any given model is done through bisection over $\beta \in [0, 10]$, halving the interval at each step until $\left|\overline{d^2} - n_{\mathrm{dim}}\right| < 0.1$. For our model with 3 degrees of freedom (DoF), $n_{\mathrm{dim}} = 3$. This is done under the assumption that the error of a calibrated pose estimator is χ^2 distributed. with $\hat{\boldsymbol{\mu}}_k$ as the position of the maximum in the search space and $\mathbf{x}_{\mathrm{GT},k}$ as the ground truth pose. This search method results in $\beta = 1.60$ with $\overline{d^2} = 3.01$ for the MulRan dataset. For the Dortmund dataset, the search results in $\beta = 1.17$ with $\overline{d^2} = 3.01$.

2) CALIBRATING THE FILTER PARAMETERS

As already presented in sections III-B, III-C3 and III-C4 the formulation of the measurement likelihood and measurement model has been motivated by the error distribution of our localization frontend. Here, the relevant parameters that need to be provided are the detection probability p_{d}, the clutter rate $\bar{\lambda}_{\mathrm{c}}$ as well as the longitudinal, lateral and orientation bias denoted by b_{lon}, b_{lat} and b_ψ. In the following, we will briefly describe how to determine reasonable estimates of these parameters based on the joint longitudinal, lateral and yaw error distribution from the per-frame localization experiment (cf. Sect. III-B). The goal is to find the portion of measurements that actually follows a Gaussian distribution (cf. Fig. 4) which yields an estimate for the detection probability p_{d}. The mean of this Gaussian distribution then provides an estimate for the pose measurement bias. Furthermore, to prevent the estimated covariance $\mathbf{R}_k$ (cf. (4)) being too overconfident, we additionally assure a minimum longitudinal, lateral and yaw variance for $\mathbf{R}_k$ denoted by $\sigma^2_{\mathrm{lon,min}}$, $\sigma^2_{\mathrm{lat,min}}$ and $\sigma^2_{\psi,\mathrm{min}}$ based on the determined Gaussian error distribution.

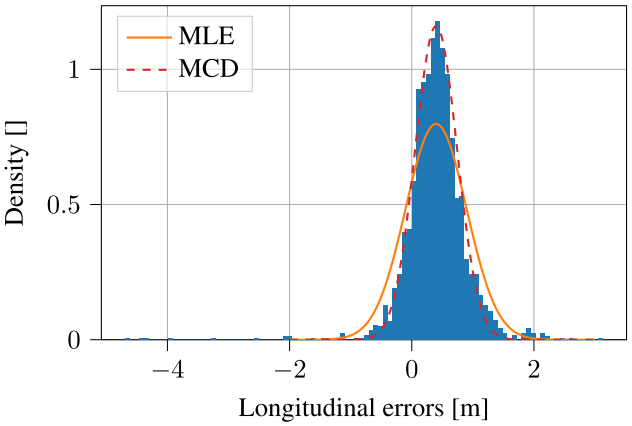


FIGURE 6. Histogram of the longitudinal errors from the offline per-frame localization experiment on the validation sequence depicted together with the Gaussian pdf of the errors estimated with maximum likelihood estimation (MLE) as well as the minimum covariance determinant (MCD) method.

A common approach to determine outliers for a given multivariate Gaussian distribution is to flag all data points as outliers whose probability density falls below a predefined threshold. In our case, a data point $\mathbf{y}_i$ of dimension n_{dim} is flagged as an outlier if the squared Mahalanobis distance

$$\left(\mathbf{y}_i - \hat{\boldsymbol{\mu}}\right)^\intercal \hat{\boldsymbol{\Sigma}} \left(\mathbf{y}_i - \hat{\boldsymbol{\mu}}\right) > \tau_{\mathrm{outlier}} \tag{34}$$

exceeds a threshold τ_{outlier} whereas $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the mean and covariance of the multivariate Gaussian respectively. The threshold τ_{outlier} is usually set based on a chi-square distribution $\chi^2_{n_{\mathrm{dim}},1-\alpha}$ with n_{dim} DoF. A common choice for the critical α level is 0.025 [90]. The problem is that the true parameters of the error distribution are usually not known such that they need to be estimated. A straightforward approach to estimate $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ is to calculate the sample mean and covariance which corresponds to maximum likelihood estimation (MLE). The problem with MLE however is that outliers contained in the data distort the estimated parameters such that outliers would appear less anomalous than they really are. To minimize the influence of outliers, the minimum covariance determinant (MCD) approach [90], [91], [92] is a robust method to estimate the mean and covariance by aiming to find the covariance matrix that best represents the dataset while minimizing its determinant. Fig. 6 shows the error distribution of Fig. 4 overlaid with corresponding Gaussian distributions fitted by the MLE and MCD method. It highlights that the MCD fit represents the error distribution more reasonably well. We apply the MCD to the joint distribution of the longitudinal, lateral and yaw errors and additionally compute the ratios of outliers and inliers. The resulting estimates of the parameters for the MulRan and Dortmund datasets are presented in Table 3 together with other relevant parameters that have been determined empirically based on the validation set data. Note, that depending on the dataset, there are different motion noise parameter since we use an odometry motion model for the MulRan dataset and a CTRV model for the Dortmund dataset.

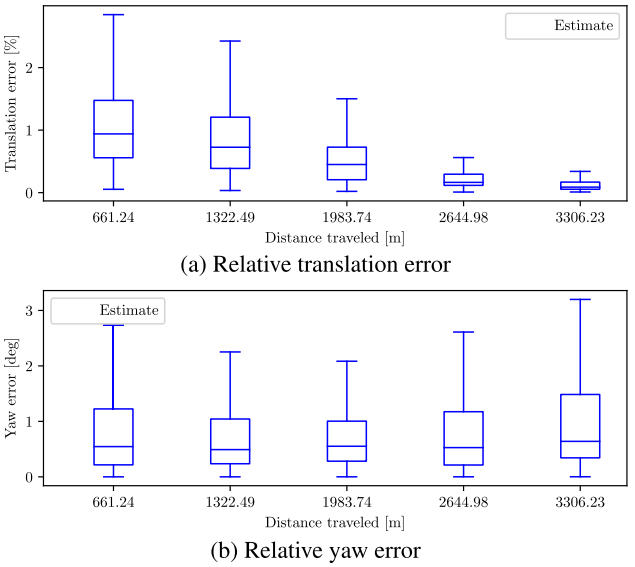
TABLE 3. Parameter settings of our proposed method for the MulRan and Dortmund dataset experiments.

Parameter	Unit	MulRan	Dortmund
p_{d}	1	0.89	0.54
λ_{c}	1	0.11	0.46
b_{lon}	m	0.40	0.00
$\sigma_{\mathrm{lon,min}}$	m	0.34	1.58
b_{lat}	m	−0.48	0.00
$\sigma_{\mathrm{lat,min}}$	m	0.17	0.34
b_{ψ}	°	0.20	0.00
$\sigma_{\psi,\mathrm{min}}$	°	1.12	1.02
$\sigma_{\delta_{\mathrm{lon}}}, \sigma_{\delta_{\mathrm{lat}}} \mid \sigma_v$	$\mathrm{m} \mid \mathrm{m\,s^{-1}}$	0.2	0.6
$\sigma_{\delta_\psi} \mid \sigma_\omega$	$^\circ \mid {}^\circ\,\mathrm{s^{-1}}$	1.0	0.5
τ_{merge}	1	1	
τ_{prune}	1	1e-6	

D. RADAR-TO-LIDAR POSE TRACKING EVALUATION ON THE MulRan DATASET

To evaluate the radar-to-lidar pose tracking performance of our approach, we evaluate the global trajectory errors as well as the relative error by travelled distance and compare the performance with a state-of-the-art approach for radar-to-lidar localization, termed RaLL [2]. Due to the aforementioned inconsistency problems with the MulRan dataset (cf. Sect. IV-A), the only sequences that are reasonable to compare are KAIST02, DCC01 and Riverside02 since here the mapping drive and the localization drive are from the same sequence such that the ground truth is valid. To provide a reasonable comparison with RaLL, we employ the single-hypothesis configuration of our approach, i.e $n_{\max} = 1$. Note, this configuration effectively corresponds to an EKF with a corresponding handling of clutter measurements and misdetections as detailed in Sect. III-C3. The employment of an EKF for pose tracking in cross-modality localization has likewise been proposed by [27], [41], and [55] as well. Equally to [41], the yaw range is divided into 21 cells with decreasing cell sizes for values closer to 0. As another baseline, we additionally evaluate a histogram filter (HF) implementation of our approach similar to [17]. Here, no Gaussian approximation of the correlation surface is performed, but the correlation surface is directly treated to be proportional to the measurement likelihood $p\left(Z_k|\mathbf{x}_k\right)$ (cf. (5)). The state space dimensions correspond to the search space of the correlation, i.e. 5 m × 5 m × 30° whereas the yaw cell spacing is linear in this case.

In addition to metric localization errors, we additionally compute the failure rate as proposed in [17]. A "failure" occurs when the estimated pose exceeds a distance threshold to the true pose. We set the threshold to 3.5 m which approximately corresponds to the distance of the farthermost point in our translational search range that is barely detectable. When a failure occurs we reset the current estimate to the true pose as our approach cannot recover from the failure when the true pose does not lie in the search range anymore. Since the localization accuracy after a reset would be unreasonably accurate, we ignore the first 5 s after a reset within the error calculation. Table 4 summarizes the quantitative results of RaLL and our approach for the KAIST02, DCC01 and Riverside02 sequence.



(a) Relative translation error

(b) Relative yaw error

FIGURE 7. Relative errors of our proposed approach on the Riverside02 sequence evaluated using the RPG Trajectory Evaluation toolbox [93].

The results show that using our proposed sensor-to-map matching model together with the adjusted training strategy (cf. Sect. III-A) and data selection (cf. Sect. IV-A), it is possible to significantly reduce the translational median error and RMSE with respect to RaLL achieving a decimeter-level localization accuracy as well as a slightly more accurate yaw estimation for both the HF and the proposed EKF approach. Our proposed EKF approach outperforms the HF in localization performance whereas the HF additionally suffers from some failures in the DCC01 and Riverside02 sequences. The reason is that, compared to the proposed EKF-based approach, the HF lacks any form of outlier rejection such that erroneous measurements can quickly move the pose estimate, i.e. the maximum of the HF state tensor, outside the search range such that the true pose is not recoverable anymore and the HF estimate diverges.

Fig. 2 visualizes input, intermediate and output data of our localization pipeline for an example frame of the KAIST02 sequence. It is visible that our localization frontend encodes widespread features from the high-resolution lidar point cloud map in the embedding. The embedding of the corresponding radar image comprises features of similar size. The resulting correlation surface depicts a nearly unimodal distribution which is approximated by a multivariate Gaussian that yields the measurement. Similar to [2], we further evaluate the relative localization errors for the Riverside02 sequence using the RPG trajectory evaluation toolbox [93]. Fig. 7 depicts the relative translation and yaw errors of our method for various traveled distances. Table 5 displays the mean relative errors compared with RaLL [2]. Compared with RaLL, our localization approach likewise results in significantly lower relative errors especially when considering maximum errors. We additionally outperform RaLL in the mean relative translation and yaw error achieving errors below 1 % and 1° respectively.

TABLE 4. Results of the radar-to-lidar pose tracking experiments on selected sequences of the MulRan dataset. Bold numbers highlight the best result for the corresponding metric.

Method	Metric		KAIST02	DCC01	Riverside02
RaLL [2]	Translation error [m]	Median	1.05	1.38	2.15
		RMSE	1.30	2.11	2.52
	Absolute yaw error [°]	Median	0.83	0.97	1.05
		RMSE	1.71	1.97	1.93
	Failure rate [%]		**0.00**	**0.00**	**0.00**
Ours (HF)	Translation error [m]	Median	0.90	0.71	0.66
		RMSE	1.09	0.82	0.87
	Absolute yaw error [°]	Median	0.79	0.66	0.68
		RMSE	1.43	**1.77**	1.66
	Failure rate [%]		**0.00**	0.05	0.22
Ours (EKF, proposed)	Translation error [m]	Median	**0.36**	**0.36**	**0.37**
		RMSE	**0.58**	**0.63**	**0.55**
	Absolute yaw error [°]	Median	**0.62**	**0.56**	**0.38**
		RMSE	**1.11**	2.23	**0.87**
	Failure rate [%]		**0.00**	**0.00**	**0.00**

TABLE 5. Average relative translation and yaw errors on the Riverside02 sequence using the toolbox from [93]. Bold numbers highlight the best result for the corresponding metric.

Metric	RaLL [2]	Ours
Mean rel. trans. errors [%]	1.08	**0.62**
Mean rel. yaw errors [°]	1.98	**0.90**

E. RADAR-TO-LIDAR PLACE RECOGNITION BASED INITIALIZATION EVALUATION

The next experiment is designed to evaluate the initialization performance of our proposed multi-hypothesis cross-modality localization method in conjunction with a state-of-the-art cross-modal PR algorithm [1] for global localization. We want to highlight at this point once again that in order to develop a true GNSS-free localization system, it is crucial that both global localization and metric pose tracking are running together in a complete localization pipeline. Therefore, assuming that the pose tracking algorithm was correctly initialized by PR, for example as done by [35], is not feasible and will eventually lead to localization failures where the reported pose may lie far off the true pose. In the following, we will describe the evaluated configurations and considered scenarios.

1) PLACE RECOGNITION METHOD

For this experiment, we use a state-of-the-art cross-modality radar-to-lidar PR method by [1]. This work builds upon Scan Context [94] which is an egocentric spatial descriptor suited for range sensor based PR [3]. The authors of [95] extended Scan Context to an end-to-end learnable rotation-invariant descriptor, called DISCO, by transforming descriptor features into the frequency domain. This allows for an additional estimation of the global orientation next to the retrieved location. Caselitz et al. [1] facilitated cross-modality PR with this descriptor by performing a joint training of all possible combinations of both single-sensor as well as radar and lidar sensor configurations using a triplet loss. Although, [1] claimed, based on their experiments, that a siamese network structure outperforms a separate encoder-decoder architecture in terms of recall, we observed in our experiments the opposite, namely that a seperate encoder-decoder structure showed improved results. Possible explanations may be that the approach requires to extract a slice from the 3D lidar point cloud near the height of the radar sensor whose dimensions however are not explicitly specified. Another reason might be the generation of the training data and labels which is unspecified as well. Nevertheless, we want to highlight, that we do not focus on the development of a novel PR method in our contribution, but merely want to achieve a realistic initialization using a given state-of-the-art approach.

For training, we employ the Adam optimizer [72]. We use a decaying learning rate scheduler with fixed step size and hyperparameters $k_0 = 1 \times 10^{-4}$, $\gamma = 0.85$ and $l = 1024$ [96]. Additionally, we use a weight decay of 0.001 and a dropout layer with rate of 0.5 in between the convolutional encoding and decoding layers of the employed U-Net architecture [97]. Similar to [1], we employ the triplet loss for training with all radar and lidar combinations for the anchor, positive and negative sample. Moreover, we choose different samples and corresponding labels for each training epoch. After training, we select the model with the smallest loss on the validation set.

Similar to [95], we save the 32^2-dimensional low-frequency signature outputs of the model in a database for the mapping drive. For a given query scan, we search for the top-n nearest neighbors based on the Euclidean distance of the signatures using the FAISS library [98]. Fig. 8 shows the top-n recall of our trained model on the validation set whereas a query is counted as a true positive if the retrieved pose is within the search range of 2.5 m and $\pm 15°$ to the true pose. Finally, the estimation of the global orientation is achieved by maximizing the cross-correlation of the retrieved frequency signatures [95].

2) TESTED CONFIGURATIONS

We want to evaluate different configurations of our proposed localization method to analyze the effects of estimating

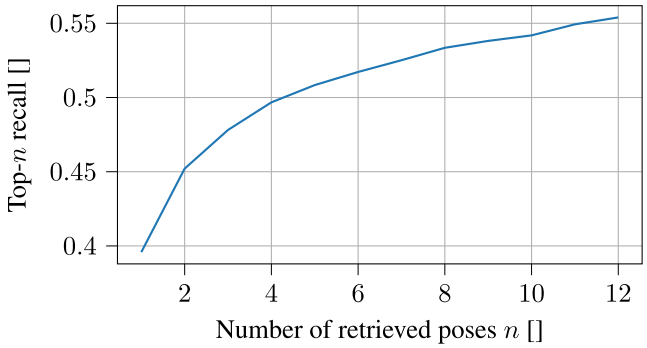


FIGURE 8. Top-*n* recall of our trained PR model based on [1], [95] on the validation set.

multiple hypotheses on the one hand and additional estimating a null hypothesis on the other hand with respect to the initialization performance. Therefore, we compare the single-hypothesis approach ($n_{\max} = 1$) against the multi-hypothesis approach where we choose $n_{\max} = 4$ which constitutes the GPU capacity in our case. For the multi-hypothesis approaches, we can additionally enable the estimation of the null hypothesis probability using either of both null hypothesis initialization strategies (NHIS) detailed in Sect. III-C7. Based on [47], we label the configurations using a NHIS with MHL-G and MHL-C for the *greedy* (G) and *conservative* (C) strategy respectively. The configuration without NHIS will be termed GSF. The resulting localization method is available if only one hypothesis is tracked and all other possible hypotheses have been pruned or merged, i.e. $n_k = 1$, and, if enabled, the null hypothesis probability is below the threshold $\tau_{\text{init}} = 0.01$. Again, we use the HF configuration as another baseline as well.

3) INITIALIZATION SCENARIOS

Similar to [2], we choose the sequences KAIST02, DCC01 and Riverside02 of the MulRan dataset for testing. To focus the evaluation on the initialization performance, we consider multiple short snippets along each sequence with different start frames where we initialize the localization based on the corresponding top-*n* retrieved poses from the PR. That is for the EKF and HF configuration, we initialize with the top-1 pose and for the GSF and MHL variants with the top-4 poses. Every snippet has a length of 100 frames, i.e. 25 s. Based on the retrieved poses from the PR, we differentiate between the following scenarios:

- *Top-1 in range*: The top-1 pose at the corresponding start frame lies within the search range of the true pose. This scenario can be seen as the baseline scenario for approaches like [27], [35], and [41] that assume, and in fact require, that the top-1 pose lies in the search range for the subsequent local pose tracking.
- *Top-1 not in range, but top-n*: in this case, the top-1 pose at the start frame is not within the search range while we enforce it to be at least 20 m away from the true pose. However, at least one of the other initial top-*n* poses lies within the search range. With this scenario, we want to evaluate the capabilities of our proposed method to identify the correct pose from the *n* initialized poses.
- *None in range*: these scenarios define the cases where none of the initially retrieved *n* poses at the start frame are close to the true pose, i.e. at least 20 m distant. Here, we want to explicitly evaluate both presented initialization strategies that additionally estimate the probability of the null hypothesis by consecutively performing PR until this probability falls below a threshold. In fact, this scenario would be the default scenario that has to be assumed when initializing the localization system with a PR method.

For every scenario we randomly select 150, 50 and 100 corresponding start frames along the KAIST02, DCC01 and Riverside02 sequence respectively.

4) INITIALIZATION-SPECIFIC EVALUATION METRICS

As additional metrics especially designated for the quantification of the initialization performance, we determine the number of initialization failures that occurred in the experiments. Initialization failures can be two-fold. Either the true pose does not lie inside the search range after the initialization time of 25 s although the localization system is available or the localization system does not become available at all within the initialization time. The latter condition may only be applied for multiple hypotheses approaches, i.e. GSF and MHL, which are unavailable if there is more than one estimated hypothesis or the null hypothesis probability is above the threshold. Based on both types of initialization failures, we can differentiate between *undetected* (Undet.) and *detected* (Det.) failures. An undetected failure corresponds to the situation where the localization system is available, but the true pose lies outside the search range after initialization. A detected failure corresponds to the case where the localization system was unavailable for the complete 100 frames.

Similar to [60], we further determine the mean and standard deviation of the convergence time of the corresponding methods, i.e. the number of seconds until the localization system becomes available for the first time. The EKF approach ($n_{\max} = 1$) as well as the HF are always deemed available with a required initialization time of 0 s as no self-assessment is performed here.

5) QUANTITATIVE RESULTS

The results of the initialization experiments for KAIST02, DCC01 and Riverside02 are summarized in Table 6. It shows that all configurations enable a successful initialization in most cases when the top-1 pose lies within the search range of the frontend with medians in the translational and absolute yaw error of < 0.4 m and $< 0.7°$ respectively. However, the approaches that do not use a NHIS, i.e. the HF, EKF and GSF, produce undetected failures. For the EKF and GSF this corresponds to situations where only one hypothesis was initialized within the search range of the true pose while

the initial yaw orientation was $> 10°$ off. Subsequently, inaccurate pose measurements were received from the frontend model such that the most likely drifted out of the search range of the true pose. For the failure of the GSF, this resulted in the problem that one of the previously less likely hypothesis became the most probable such that the estimation falsely converged to this wrong hypothesis. Again, the HF failed due to clutter measurements mostly occurring in the Riverside02 sequence (cf. Sect. IV-D) such that the true pose moved out of the search range. The MHL approaches do not suffer from the aforementioned problems as they reinitialize multiple new hypotheses during the initialization phase such that it becomes very likely that more than one hypothesis will be initialized close to the true pose which likewise leads to multiple pose measurements in the search range. Here, the estimation eventually converges around the true pose. This reveals that employing a NHIS robustifies the initialization even in the case when the top-1 pose is already close to the true pose. The drawback is that the convergence time increases when using multiple hypotheses (cf. Table 6). This is expected as it requires some update steps until the weights of the incorrect hypotheses fall below the pruning threshold and the null hypothesis probability finally falls below τ_{init}.

The results of the scenarios where the top-1 pose is not in range highlight that using the top-1 PR pose for initialization without any method for validation will lead to localization failures with arbitrarily large errors that may even exceed several hundreds of meters. However, if at least one of the top-n poses lies within the search range, the correct pose can be identified with 0 % undetected failures when using one of the proposed MHL approaches. Again, the GSF shows a less robust performance where it converged to a wrong hypothesis in some cases.

Finally, when *none* of the initially retrieved poses lies close to the true pose, only the MHL approaches may succeed while the other approaches fail for all experiments. Here, the GSF approach may detect some failures in cases where the ambiguities between the four initialized hypotheses could not be resolved within the initialization phase such that more than one hypothesis survived after 100 frames. The approach therefore already provides some capabilities of detecting ambiguous situations during initialization. However, in most cases one of the four tracked, but incorrect, hypotheses still becomes much more likely with respect to the other hypotheses such that all other hypotheses are pruned. The greedy MHL method generally achieves faster convergence times and fewer failures when considering detected failures. This is due to the property that the conservative method requires more time for initialization as all hypotheses except one have to be pruned before new hypotheses can be initialized. Since it may happen that ambiguities cannot be resolved over the whole initialization sequence, as visible in the GSF results, the resulting rate of detected failures is much higher for the conservative initialization method. The faster initialization time of the greedy method however comes with the cost of an inferior robustness compared to MHL-C. This may even lead to undetected localization failures for very ambiguous situations which are present in the Riverside02 sequence. Overall, the MHL-C method shows the most robust initialization performance and likewise achieves the lowest metric localization errors when initialized.

6) QUALITATIVE RESULTS

Fig. 9 depicts an example initialization experiment using the MHL-G config on the KAIST02 sequence where none of the initially retrieved PR poses lied within the search range. Fig. 9a shows the most likely hypothesis (top row) as well as the currently closest hypothesis to the true pose (bottom row) depicted by a blue triangle and 2σ covariance ellipse overlaid on the corresponding map tiles for different frames. The middle row shows the radar sensor image of the corresponding frame. The red triangles and ellipses represent the pose measurements and associated uncertainties that resulted from the Gaussian approximation of the correlation surface. Finally, the green triangles depict the true poses. The edge colors of the images correspond to different frames that are accordingly marked with vertical lines of the same color in the plot of the hypotheses log weights below in Fig. 9b. At the beginning, the initialized hypotheses have similar weights whereas none of the hypotheses is close to the true pose. Over the next subsequent frames, odometry and pose measurements are fused within the prediction and update step of the MHL-G filter, which results in relative changes between the hypotheses' weights. However, the null hypothesis probability stays constant as no hypotheses gets pruned or merged. After frame 14 the situation changes. Here, the weight of the least likely hypothesis falls below τ_{prune} such that it gets pruned and the PR is retriggered. This process repeats several times until it happens that at frame 52 the PR actually returns a pose that is within the search range of the true pose which is visible in the mid-bottom image in Fig. 9a. The most likely hypothesis at that frame is by this time far off and the estimation even moved outside the driveable area due to inconsistencies between the odometry and pose measurements. In the subsequent frames, the MHL-G algorithm correctly and quickly identifies that the newly initialized hypothesis is much more likely than the previous most likely hypothesis such that already two frames later a hypothesis switch takes place and the newly initialized hypothesis becomes the most likely hypothesis. Afterwards, this hypothesis consistently stays the most likely hypothesis while the null hypothesis monotonically decreases whenever another hypothesis is pruned or merged until $\hat{w}_k^{(0)}$ falls below τ_{init}.

F. LIDAR-TO-AERIAL-IMAGERY POSE TRACKING EVALUATION

To show that our approach is suited for different sensor modalities, we performed lidar-to-aerial-imagery localization, also known as geo-tracking [41], on a self-recorded

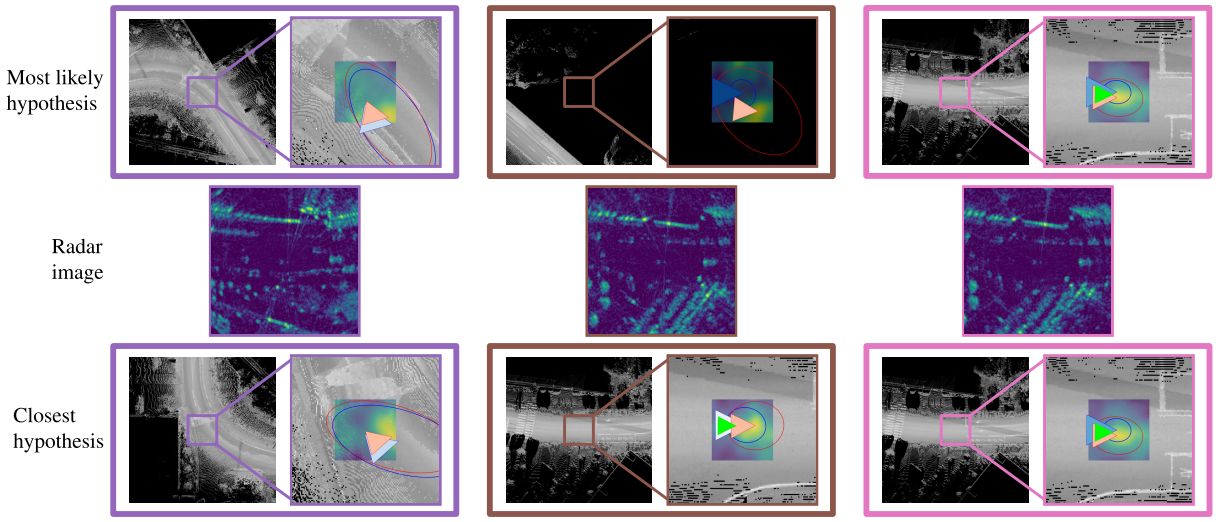
TABLE 6. Results of the radar-to-lidar initialization experiments on selected sequences of the MulRan dataset. Bold numbers highlight the best result for the corresponding metric.

Sequence	Scenario	Config.	Failures [%]		Convergence time [s]		Trans. error [m]		Abs. yaw error [°]	
			Undet.	Det.	Mean	Std.	Median	RMSE	Median	RMSE
KAIST02 (150 runs)	Top-1 in range	HF	**0.00**	0.00	**0.00**	**0.00**	0.88	1.09	0.83	1.48
		EKF	0.67	0.00	**0.00**	**0.00**	0.36	2.30	0.67	1.88
		GSF	**0.00**	0.67	1.64	0.72	0.35	**0.54**	0.66	1.12
		MHL-G	**0.00**	0.00	4.76	1.13	0.34	**0.54**	0.64	1.08
		MHL-C	**0.00**	0.00	8.59	2.19	**0.33**	**0.54**	**0.63**	**1.04**
	Top-1 not in range, but top-n	HF	100.00	0.00	**0.00**	**0.00**	348.11	418.63	95.71	107.45
		EKF	100.00	0.00	**0.00**	**0.00**	349.25	418.57	91.39	106.09
		GSF	1.33	0.00	2.21	1.72	0.35	39.56	**0.59**	9.70
		MHL-G	**0.00**	0.00	5.44	1.51	0.35	**0.51**	0.60	**1.03**
		MHL-C	**0.00**	0.67	9.17	2.24	**0.34**	**0.51**	**0.59**	1.05
	None in range	HF	100.00	0.00	**0.00**	**0.00**	395.87	434.32	92.74	103.01
		EKF	100.00	0.00	**0.00**	**0.00**	400.36	436.76	95.12	101.73
		GSF	81.33	18.67	13.95	6.24	402.47	460.06	103.34	110.51
		MHL-G	**0.00**	9.33	13.99	5.51	0.38	0.55	0.67	1.26
		MHL-C	**0.00**	64.00	19.46	3.89	**0.35**	**0.51**	**0.61**	**1.03**
DCC01 (50 runs)	Top-1 in range	HF	**0.00**	0.00	**0.00**	**0.00**	0.69	0.77	0.67	1.71
		EKF	**0.00**	0.00	**0.00**	**0.00**	0.37	0.52	0.54	1.02
		GSF	**0.00**	0.00	1.45	0.81	0.37	0.49	0.53	1.03
		MHL-G	**0.00**	0.00	4.61	1.23	**0.36**	0.49	0.51	0.94
		MHL-C	**0.00**	0.00	7.22	2.60	**0.36**	**0.48**	**0.49**	**0.90**
	Top-1 not in range, but top-n	HF	100.00	0.00	**0.00**	**0.00**	300.59	332.49	107.60	110.39
		EKF	100.00	0.00	**0.00**	**0.00**	292.49	334.05	96.84	109.60
		GSF	2.00	0.00	2.26	1.39	**0.34**	2.79	0.58	1.79
		MHL-G	**0.00**	0.00	5.22	1.74	0.35	0.43	0.58	1.03
		MHL-C	**0.00**	0.00	8.99	3.41	**0.34**	**0.42**	**0.54**	**0.97**
	None in range	HF	100.00	0.00	**0.00**	**0.00**	257.38	328.50	96.60	105.84
		EKF	100.00	0.00	**0.00**	**0.00**	242.12	322.48	92.77	104.40
		GSF	90.00	10.00	12.27	6.11	312.84	357.54	112.73	121.08
		MHL-G	**0.00**	18.00	10.87	4.22	0.35	**0.44**	0.44	1.03
		MHL-C	**0.00**	52.00	16.94	5.39	**0.34**	0.45	**0.43**	**1.00**
Riverside02 (100 runs)	Top-1 in range	HF	4.00	0.00	**0.00**	**0.00**	0.64	0.98	0.70	1.71
		EKF	**0.00**	0.00	**0.00**	**0.00**	**0.36**	0.56	0.35	0.78
		GSF	1.00	0.00	2.47	2.04	**0.36**	53.32	0.34	12.87
		MHL-G	**0.00**	0.00	7.58	3.68	**0.36**	**0.44**	**0.32**	**0.67**
		MHL-C	**0.00**	3.00	12.67	4.38	0.37	0.46	0.33	0.69
	Top-1 not in range, but top-n	HF	100.00	0.00	**0.00**	**0.00**	453.62	661.42	45.37	73.62
		EKF	100.00	0.00	**0.00**	**0.00**	366.09	605.62	42.55	70.19
		GSF	3.00	0.00	3.67	2.75	**0.36**	100.31	0.36	7.71
		MHL-G	**0.00**	0.00	8.21	3.38	**0.36**	0.46	0.35	**0.67**
		MHL-C	**0.00**	6.00	15.33	5.37	**0.36**	**0.44**	**0.34**	0.69
	None in range	HF	100.00	0.00	**0.00**	**0.00**	382.39	747.28	49.43	78.67
		EKF	100.00	0.00	**0.00**	**0.00**	370.93	738.95	47.78	80.26
		GSF	88.00	12.00	9.72	5.30	268.32	670.20	26.42	67.49
		MHL-G	15.00	15.00	14.83	4.95	0.40	198.69	0.47	14.63
		MHL-C	**0.00**	71.00	19.03	4.04	**0.35**	**0.41**	**0.35**	**0.84**

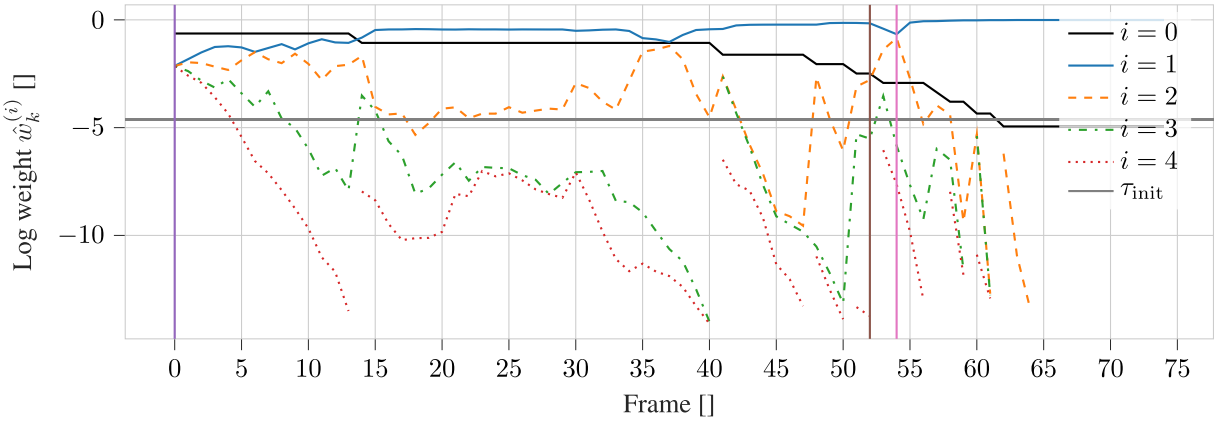
TABLE 7. Results of our proposed EKF configuration for the lidar-to-aerial-imagery pose tracking experiments on a highway and urban sequence of the Dortmund dataset.

Sequence	Trans. error [m]		Abs. yaw error [°]		Failure rate [%]
	Mean	SD	Mean	SD	
Highway	0.96	0.68	0.77	0.32	0.06
Urban	0.85	0.43	0.57	0.42	0.00

dataset around the city of Dortmund, Germany, for the EKF configuration initialized with the GT. Table 7 reports the mean and standard deviation of the translation and yaw error for both scenarios. With an average translation error < 1 m for both a highway and an urban scenario, we achieve a similar localization accuracy with respect to the state-of-the-art geo-tracking approach of [41]. However, there are some failures that occurred in the highway experiment. Fig. 10 depicts two situations of the highway sequence. The top row depicts a scenario where the localization is successful whereas the bottom row shows a situation where a failure occurred. The cause of the failure is visible when regarding the correlation surface. In case of a successful localization, a clear smooth peak is visible in the correlation surface. On the other hand, in case of a failure the correlation distribution is clearly multimodal such that it can only be poorly approximated by a Gaussian distribution.



(a) Radar image as well as most likely and closest hypothesis to true pose overlaid over the corresponding map locations at selected frames.



(b) Evolution of the estimated log weights of the tracked hypotheses including the null hypothesis.

FIGURE 9. Evolution of an initialization scenario on the KAIST02 sequence of the MulRan dataset [3] using the MHL-G configuration where initially none of the tracked hypotheses is close to the true pose. Subfigure (a) shows the most likely (top row) and currently closest hypothesis to the GT (bottom row) after the update step together with the corresponding radar sensor image (middle row). The *blue triangles with covariance ellipses* correspond to the estimated hypotheses whereas the *red* counterparts depict the current pose observations. The *green triangle* corresponds to the true pose. The colors of the image borders correspond to specific time frames equally highlighted in subfigure (b) which depicts the corresponding evolution of the hypotheses log weights during the experiment together with the null hypothesis probability threshold τ_{init}.

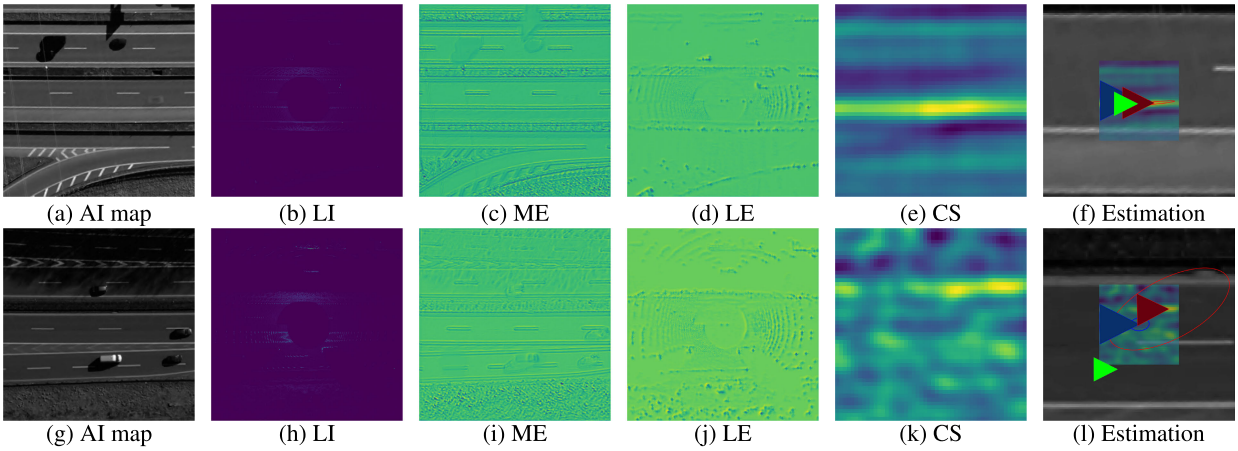


FIGURE 10. Depiction of input, intermediate and output data of the proposed localization approach for two frames of the highway sequence. Subfigures (a)-(f) represent a scenario where the localization was successful. Subfigures (g)-(l) shows a situation where a failure occurred. The subcaptions comprise the following abbreviations: AI: aerial image, LI: lidar image, ME: map embedding, LE: lidar embedding, CS: correlation surface.

V. CONCLUSION

In this work, we presented a true GNSS-free cross-modality global localization method by combining place recognition with multi-hypothesis pose tracking to validate if one of the retrieved poses from PR actually corresponds to the true pose. Here, the multi-hypothesis pose tracking algorithm obtains measurements from our deep sensor-to-map matching network that is capable of extracting relevant features of both two-dimensional and three-dimensional map and sensor data of different modalities. Together with the latest odometry measurements it is possible to infer the correct hypothesis over time. The multi-hypothesis tracking algorithm is realized by a Gaussian sum filter which allows for an initialization with the top-n retrieved poses from the PR. Equivalent multi-hypothesis solutions based on commonly used histogram or particle filters would be computationally hardly manageable. Since the top-n recall of current PR methods is below 100 %, we additionally account for the case that none of the initialized poses is close to the true pose by introducing the concept of a null hypothesis and estimate its probability during the initialization phase. Our localization method continuously eliminates unlikely hypotheses while lowering the null hypothesis probability through the initialization of new hypotheses based on the PR until one likely pose hypothesis remains and the null hypothesis probability falls below a threshold. Only starting from this point, the localization system becomes available. In extensive experiments, we show that our system achieves 0 % undetected initialization failures on the MulRan dataset while achieving state-of-the-art metric cross-modality localization accuracy on both a radar-to-lidar as well as a lidar-to-aerial-imagery dataset. Future work will focus on the extension of the approach to camera images as sensor input data enabling cross-view geo-localization (CVGL) [41]. Further enhancements will focus on localization robustness and accuracy.

ACKNOWLEDGMENT

We acknowledge financial support by Deutsche Forschungsgemeinschaft and Technische Universitat Dortmund/TU Dortmund University within the funding programme Open Access Costs.

REFERENCES

[1] H. Yin, X. Xu, Y. Wang, and R. Xiong, "Radar-to-LiDAR: Heterogeneous place recognition via joint learning," *Frontiers Robot. AI*, vol. 8, May 2021, Art. no. 661199.

[2] H. Yin, R. Chen, Y. Wang, and R. Xiong, "RaLL: End-to-end radar localization on LiDAR map using differentiable measurement model," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6737–6750, Jul. 2022.

[3] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "MulRan: Multimodal range dataset for urban place recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Paris, France, May 2020, pp. 6246–6253.

[4] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, and A. F. De Souza, "Self-driving cars: A survey," *Expert Syst. Appl.*, vol. 165, Feb. 2021, Art. no. 113816.

[5] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge, MA, USA: MIT Press, 2005.

[6] W. Ma, I. Tartavull, I. A. Bârsan, S. Wang, M. Bai, G. Mattyus, N. Homayounfar, S. K. Lakshmikanth, A. Pokrovsky, and R. Urtasun, "Exploiting sparse semantic HD maps for self-driving vehicle localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 5304–5311.

[7] M. Karaim, M. Elsheikh, A. Noureldin, and R. Rustamov, "GNSS error sources," in *Multifunctional Operation and Application of GPS*, 1st ed., May 2018, pp. 69–85. [Online]. Available: https://directory.doabooks.org/handle/20.500.12854/54063?show=full and https://www.intechopen.com/books/6540, doi: 10.5772/intechopen.71221.

[8] M. Harris, "Military tests that jam and spoof GPS signals are an accident waiting to happen," *IEEE Spectr.*, vol. 58, no. 2, pp. 22–27, Feb. 2021.

[9] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[10] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19516–19547, 2021.

[11] P. Yin, S. Zhao, I. Cisneros, A. Abuduweili, G. Huang, M. Milford, C. Liu, H. Choset, and S. Scherer, "General place recognition survey: Towards the real-world autonomy age," 2022, *arXiv:2209.04497*.

[12] J. Levinson, M. Montemerlo, and S. Thrun, "Map-based precision vehicle localization in urban environments," in *Robotics: Science and Systems*, vol. 4. Cambridge, MA, USA: MIT Press, 2007, p. 1.

[13] J. Levinson and S. Thrun, "Robust vehicle localization in urban environments using probabilistic maps," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 4372–4378.

[14] A. Y. Hata and D. F. Wolf, "Feature detection for vehicle localization in urban environments using a multilayer LiDAR," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 420–429, Feb. 2016.

[15] R. W. Wolcott and R. M. Eustice, "Robust LiDAR localization using multiresolution Gaussian mixture maps for autonomous driving," *Int. J. Robot. Res.*, vol. 36, no. 3, pp. 292–319, Mar. 2017.

[16] G. Wan, X. Yang, R. Cai, H. Li, Y. Zhou, H. Wang, and S. Song, "Robust and precise vehicle localization based on multi-sensor fusion in diverse city scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4670–4677.

[17] I. A. Bârsan, S. Wang, A. Pokrovsky, and R. Urtasun, "Learning to localize using a LiDAR intensity map," in *Proc. 2nd Conf. Robot Learn.*, in Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87, Oct. 2018, pp. 605–616.

[18] Z. Tao, P. Bonnifait, V. Frémont, and J. Ibañez-Guzman, "Mapping and localization using GPS, lane markings and proprioceptive sensors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 406–412.

[19] J. Ziegler, H. Lategahn, M. Schreiber, C. G. Keller, C. Knöppel, J. Hipp, M. Haueis, and C. Stiller, "Video based localization for Bertha," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 1231–1238.

[20] R. P. D. Vivacqua, M. Bertozzi, P. Cerri, F. N. Martins, and R. F. Vassallo, "Self-localization based on visual lane marking maps: An accurate low-cost approach for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 582–597, Feb. 2018.

[21] E. Ward and J. Folkesson, "Vehicle localization with low cost radar sensors," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2016, pp. 864–870.

[22] F. Schuster, M. Wörner, C. G. Keller, M. Haueis, and C. Curio, "Robust localization based on radar signal clustering," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2016, pp. 839–844.

[23] K. Yoneda, N. Hashimoto, R. Yanase, M. Aldibaja, and N. Suganuma, "Vehicle localization using 76 GHz omnidirectional millimeter-wave radar for winter automated driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 971–977.

[24] A. Schindler, "Vehicle self-localization with high-precision digital maps," in *Proc. IEEE Intell. Vehicles Symp. Workshops (IV Workshops)*, Jun. 2013, pp. 134–139.

[25] M. Lundgren, E. Stenborg, L. Svensson, and L. Hammarstrand, "Vehicle self-localization using off-the-shelf sensors and a detailed map," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 522–528.

[26] N. Stannartz, M. Theers, M. Sons, A. Llarena, M. Kuhn, O. M. Kind, and T. Bertram, "Efficient localization on highways employing public HD maps and series-production sensors," in *Proc. 21st Internationales Stuttgarter Symp.* Wiesbaden, Germany: Springer, 2021, pp. 395–409.

[27] F. Fervers, S. Bullinger, C. Bodensteiner, M. Arens, and R. Stiefelhagen, "Uncertainty-aware vision-based metric cross-view geolocalization," 2022, *arXiv:2211.12145*.

[28] Z. Tao, P. Bonnifait, V. Frémont, J. Ibanez-Guzman, and S. Bonnet, "Road-centered map-aided localization for driverless cars using single-frequency GNSS receivers," *J. Field Robot.*, vol. 34, no. 5, pp. 1010–1033, Aug. 2017.

[29] J. Pauls, T. Strauss, C. Hasberg, M. Lauer, and C. Stiller, "HD map verification without accurate localization prior using spatio-semantic 1D signals," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 680–686.

[30] J. K. Suhr, J. Jang, D. Min, and H. G. Jung, "Sensor fusion-based low-cost vehicle localization system for complex urban environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1078–1086, May 2017.
[31] D. Wilbers, C. Merfels, and C. Stachniss, "Localization with sliding window factor graphs on third-party maps for automated driving," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5951–5957.
[32] F. Ghallabi, G. El-Haj-Shhade, M. Mittet, and F. Nashashibi, "LiDAR-based road signs detection for vehicle localization in an HD map," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1484–1490.
[33] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3D LiDAR maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 1926–1931.
[34] D. Cattaneo, M. Vaghi, S. Fontana, A. L. Ballardini, and D. G. Sorrenti, "Global visual localization in LiDAR-maps through shared 2D-3D embedding space," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4365–4371.
[35] T. Y. Tang, D. De Martini, and P. Newman, "Get to the point: Learning LiDAR place recognition and metric localisation using overhead imagery," *Proc. Robot., Sci. Syst.*, 2021, pp. 25–35.
[36] X. Yu, B. Zhou, Z. Chang, K. Qian, and F. Fang, "MMDF: Multi-modal deep feature based place recognition of mobile robots with applications on cross-scene navigation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6742–6749, Jul. 2022.
[37] H. Yin, Y. Wang, L. Tang, and R. Xiong, "Radar-on-LiDAR: Metric radar localization on prior LiDAR maps," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot. (RCAR)*, Sep. 2020, pp. 1–7.
[38] K. Burnett, Y. Wu, D. J. Yoon, A. P. Schoellig, and T. D. Barfoot, "Are we ready for radar to replace LiDAR in all-weather mapping and localization?" *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10328–10335, Oct. 2022.
[39] T. Y. Tang, D. De Martini, D. Barnes, and P. Newman, "RSL-Net: Localising in satellite images from a radar on the ground," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1087–1094, Apr. 2020.
[40] T. Y. Tang, D. De Martini, S. Wu, and P. Newman, "Self-supervised learning for using overhead imagery as maps in outdoor range sensor localization," *Int. J. Robot. Res.*, vol. 40, nos. 12–14, pp. 1488–1509, Dec. 2021.
[41] F. Fervers, S. Bullinger, C. Bodensteiner, M. Arens, and R. Stiefelhagen, "Continuous self-localization on aerial images using visual and LiDAR sensors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 7028–7035.
[42] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16989–16999.
[43] M. Stübler, "Self-assessing localization and long-term mapping using random finite sets," Ph.D. thesis, Inst. Meas., Control Microtechnol., Ulm Univ., Ulm, Germany, 2018.
[44] T. G. R. Reid, S. E. Houts, R. Cammarata, G. Mills, S. Agarwal, A. Vora, and G. Pandey, "Localization requirements for autonomous vehicles," 2019, *arXiv:1906.01061*.
[45] D. Alspach and H. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Trans. Autom. Control*, vol. AC-17, no. 4, pp. 439–448, Aug. 1972.
[46] P. Jensfelt and S. Kristensen, "Active global localization for a mobile robot using multiple hypothesis tracking," *IEEE Trans. Robot. Autom.*, vol. 17, no. 5, pp. 748–760, Oct. 2001.
[47] P. Jensfelt, "Approaches to mobile robot localization in indoor environments," Ph.D. thesis, Dept. Signals, Sensors Syst., Roy. Inst. Technol. (KTH), Stockholm, Sweden, 2001.
[48] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel CNN for efficient 3D deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 965–975.
[49] S. Hu and G. H. Lee, "Image-based geo-localization using satellite imagery," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1205–1219, May 2020.
[50] D. Kim and M. R. Walter, "Satellite image-based localization via learned embeddings," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2073–2080.
[51] J. L. Blanco-Claraco, F. Mañas-Alvarez, J. L. Torres-Moreno, F. Rodriguez, and A. Gimenez-Fernandez, "Benchmarking particle filter algorithms for efficient velodyne-based vehicle localization," *Sensors*, vol. 19, no. 14, p. 3155, Jul. 2019.
[52] R. P. Mahler, *Statistical Multisource-Multitarget Information Fusion*, vol. 685. Norwood, MA, USA: Artech House, 2007.
[53] S. Challa, M. R. Morelande, D. Mušicki, and R. J. Evans, *Fundamentals of Object Tracking*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[54] T. Griebel, J. Müller, P. Geisler, C. Hermann, M. Herrmann, M. Buchholz, and K. Dietmayer, "Self-assessment for single-object tracking in clutter using subjective logic," 2022, *arXiv:2206.07449*.
[55] A. Vora, S. Agarwal, G. Pandey, and J. McBride, "Aerial imagery based LiDAR localization for autonomous vehicles," 2020, *arXiv:2003.11192*.
[56] P. A. Iannucci, L. Narula, and T. E. Humphreys, "Cross-modal localization: Using automotive radar for absolute geolocation within a map produced with visible-light imagery," in *Proc. IEEE/ION Position, Location Navigat. Symp. (PLANS)*, Apr. 2020, pp. 285–296.
[57] E. B. Olson, "Real-time correlative scan matching," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 4387–4393.
[58] D. Barnes, R. Weston, and I. Posner, "Masking by moving: Learning distraction-free radar odometry from pose information," in *Proc. Conf. Robot Learn.*, in Proceedings of Machine Learning Research, L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100, Oct./Nov. 2020, pp. 303–316.
[59] F. Yan, O. Vysotska, and C. Stachniss, "Global localization on OpenStreetMap using 4-bit semantic descriptors," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Sep. 2019, pp. 1–7.
[60] I. D. Miller, A. Cowley, R. Konkimalla, S. S. Shivakumar, T. Nguyen, T. Smith, C. J. Taylor, and V. Kumar, "Any way you look at it: Semantic crossview localization and mapping with LiDAR," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2397–2404, Apr. 2021.
[61] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, "Deep visual geo-localization benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5386–5397.
[62] J. Rabe, M. Necker, and C. Stiller, "Ego-lane estimation for lane-level navigation in urban scenarios," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2016, pp. 896–901.
[63] J. Rabe, M. Hübner, M. Necker, and C. Stiller, "Ego-lane estimation for downtown lane-level navigation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1152–1157.
[64] F. Li, P. Bonnifait, J. Ibanez-Guzman, and C. Zinoune, "Lane-level map-matching with integrity on high-definition maps," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1176–1181.
[65] F. Li, P. Bonnifait, and J. Ibanez-Guzman, "Estimating localization uncertainty using multi-hypothesis map-matching on high-definition road maps," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 1–6.
[66] F. Li, P. Bonnifait, and J. Ibañez-Guzmán, "Map-aided dead-reckoning with lane-level maps and integrity monitoring," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 1, pp. 81–91, Mar. 2018.
[67] M. A. Brubaker, A. Geiger, and R. Urtasun, "Lost! Leveraging the crowd for probabilistic visual self-localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3057–3064.
[68] M. A. Brubaker, A. Geiger, and R. Urtasun, "Map-based probabilistic visual self-localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 652–665, Apr. 2016.
[69] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.
[70] J. Liu, S. Mills, and B. McCane, "Variational autoencoder for 3D voxel compression," in *Proc. 35th Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2020, pp. 1–6.
[71] *Digitale Orthophotos NRW*, Geobasis North Rhine-Westphalia District Government Cologne, Germany, 2010.
[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[73] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[74] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," presented at the Int. Conf. Learn. Represent. (ICLR), 2019. [Online]. Available: https://dblp.org/db/conf/iclr/iclr2019.html and https://openreview.net/group?id=ICLR.cc/2019/Conference
[75] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," 2014, *arXiv:1406.2080*.
[76] J. Sola, "Course on SLAM," Inst. Robot. Ind. Inform., CSIC-UPC, Barcelona, Spain, Tech. Rep. IRI-TR-16-04, 2017.
[77] R. Schubert, C. Adam, M. Obst, N. Mattern, V. Leonhardt, and G. Wanielik, "Empirical evaluation of vehicular models for ego motion estimation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 534–539.

[78] B. N. Vo and W. K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.
[79] F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Found. Trends Robot.*, vol. 6, nos. 1–2, pp. 1–139, 2017.
[80] J. Al Hage, P. Xu, P. Bonnifait, and J. Ibanez-Guzman, "Localization integrity for intelligent vehicles through fault detection and position error characterization," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 2978–2990, Apr. 2022.
[81] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The Oxford radar RobotCar dataset: A radar extension to the Oxford RobotCar dataset," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 6433–6438.
[82] D. Barnes and I. Posner, "Under the radar: Learning to predict robust keypoints for odometry estimation and metric localisation in radar," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 9484–9490.
[83] P. Kung, C. Wang, and W. Lin, "A normal distribution transform-based radar odometry designed for scanning and automotive radars," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 14417–14423.
[84] K. Burnett, D. J. Yoon, A. P. Schoellig, and T. D. Barfoot, "Radar odometry combining probabilistic estimation and unsupervised feature learning," 2021, *arXiv:2105.14152*.
[85] K. Burnett, A. P. Schoellig, and T. D. Barfoot, "Do we need to compensate for motion distortion and Doppler effects in spinning radar navigation?" *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 771–778, Apr. 2021.
[86] D. Adolfsson, M. Magnusson, A. Alhashimi, A. J. Lilienthal, and H. Andreasson, "LiDAR-level localization with radar? The CFEAR approach to accurate, fast, and robust large-scale radar odometry in diverse environments," *IEEE Trans. Robot.*, vol. 39, no. 2, pp. 1476–1495, Apr. 2023.
[87] T. Li, C. Jiang, Z. Bian, M. Wang, and X. Niu, "A review of true orthophoto rectification algorithms," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 780, no. 2, Mar. 2020, Art. no. 022035.
[88] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70, Aug. 2017, pp. 1321–1330.
[89] R. Kümmerle, B. Steder, C. Dornhege, M. Ruhnke, G. Grisetti, C. Stachniss, and A. Kleiner, "On measuring the accuracy of SLAM algorithms," *Auton. Robots*, vol. 27, no. 4, pp. 387–407, Nov. 2009.
[90] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, Aug. 1999.
[91] P. J. Rousseeuw, "Least median of squares regression," *J. Amer. Statist. Assoc.*, vol. 79, no. 388, pp. 871–880, 1984.
[92] R. W. Butler, P. L. Davies, and M. Jhun, "Asymptotics for the minimum covariance determinant estimator," *Ann. Statist.*, vol. 21, no. 3, pp. 1385–1400, Sep. 1993.
[93] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 7244–7251.
[94] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 4802–4809.
[95] X. Xu, H. Yin, Z. Chen, Y. Li, Y. Wang, and R. Xiong, "DiSCO: Differentiable scan context with orientation," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2791–2798, Apr. 2021.
[96] Y. Wu, L. Liu, J. Bae, K. Chow, A. Iyengar, C. Pu, W. Wei, L. Yu, and Q. Zhang, "Demystifying learning rate policies for high accuracy training of deep neural networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 1971–1980.
[97] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.
[98] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

NIKLAS STANNARTZ received the B.S. and M.S. degrees in electrical engineering and information technology from TU Dortmund University, Dortmund, Germany, in 2015 and 2018, respectively, where he is currently pursuing the Dr.-Ing. degree in electrical engineering and information technology.

Since 2018, he has been a Research Associate with the Department of Electrical Engineering and Information Technology, TU Dortmund University. His research interest includes the development of accurate and ambiguity-aware localization methods for automated vehicles using on-board perception sensors.

Mr. Stannartz is a member of the research group that won the science competition "Forum Junge Spitzenforscher" hosted by the Center for Entrepreneurship and Transfer, TU Dortmund University. He was a recipient of the Elmos Semiconductor AG Award, in 2016, and the VDE Rhein-Ruhr e. V. Award for Outstanding Graduation, in 2018.

STEFAN SCHÜTTE received the B.S. and M.S. degrees in electrical engineering and information technology from TU Dortmund University, Dortmund, Germany, in 2018 and 2021, respectively, where he is currently pursuing the Dr.-Ing. degree.

He has been a Research Assistant with the Department of Electrical Engineering and Information Technology, TU Dortmund University, since 2021. His research interest includes cross-modality localization approaches for automated vehicles.

MARKUS KUHN received the Dipl.-Phys. and Dr.rer.nat. degrees in physics from the University of Siegen, Siegen, Germany, in 2005 and 2008, respectively.

In 2004, he was a Summer Student with CERN, working on the ATLAS particle detector. From 2008 to 2011, he was a Product Development and Design Engineer with Voith Paper. From 2012 to 2017, he was a Calculation and Research Engineer with Andritz Separation, working in mechanical engineering, structural mechanics, and separation technology. In 2017, he joined the ZF Group, Düsseldorf, Germany, as an Algorithm Developer for automated driving, where he has been leading the Localization and Environmental Model Team, since 2020.

TORSTEN BERTRAM received the Dipl.-Ing. and Dr.-Ing. degrees in mechanical engineering from Gerhard Mercator University Duisburg, Duisburg, Germany, in 1990 and 1995, respectively.

In 1990, he joined the Department of Mechanical Engineering, Gerhard Mercator University Duisburg, as a Research Associate. From 1995 to 1998, he was a Subject Specialist with the Corporate Research Division, Bosch Group, Stuttgart, Germany. In 1998, he returned to Gerhard Mercator University Duisburg as an Assistant Professor. In 2002, he became a Distinguished Professor with the Department of Mechanical Engineering, TU Ilmenau University, Ilmenau, Germany. Since 2005, he has been a member of the Department of Electrical Engineering and Information Technology, TU Dortmund University, Dortmund, Germany, and also a Distinguished Professor of systems and control engineering. His research interests include control theory, artificial intelligence, and their application to mechatronics, service robotics, and automotive systems.