Semantic-Guided Geometric Feature Extraction from Dense LiDAR for Vehicle Localization with Abstract Maps

Mohamed Wahbah [1], Rozhin Moftizadeh[1], Christopher Klugmann[2], Daniel Kondermann[2], Ingo Neumann[1], Hamza Alkhatib[1]

[1] Geodetic Institute, Leibniz University Hanovver, Germany - (wahbah,moftizadeh,neumann,alkhatib)@gih.uni-hannover.de
[2] Quality Match GmbH, Heidelberg Germany - (christopher.klugmann,dk)@quality-match.com

Keywords: Feature Extraction, Geo-referencing, LoD2 models, Digital Terrain Models (DTM).

Abstract

High-precision vehicle localization in GNSS-denied urban areas requires alternatives to costly HD maps. In this paper, we present a novel framework for feature extraction and benchmark generation to enable high-precision localization using abstract LoD2/DTM maps as a replacement for HD maps. Our first contribution, a semantic-geometric pipeline, processes dense LiDAR and camera data to extract map primitives. This is accomplished by a RANSAC-fitted ground plane extraction step, followed by a semantic filter that discards dynamic objects. Finally, geometric clustering (HDBSCAN) and RANSAC plane fitting isolate large-scale vertical facades. Our second contribution, a multi-stage GT generation framework, resolves annotation ambiguity using a Human-In-The-Loop (HITL) system. A robust 2D pose is computed by finding the geometric median of bootstrapped transformation samples on the $SE(2)$ manifold, which is then refined to a 6-Degree-of-Freedom pose via point-to-plane ICP, before being validated by a human for a final check. We evaluated our feature extraction pipeline against the generated benchmark, achieving 95.04% precision and 83.74% recall. An analysis of this performance shows the pipeline correctly rejects small, ambiguous features while achieving high recall on all large, stable features, proving its suitability for a robust localization filter.

1. Introduction & Background

Accurate and robust global localization is a fundamental prerequisite for higher levels of driving automation, enabling core functions such as path planning and mission execution. The challenge is particularly critical in dense urban environments, where positional awareness requires accuracy on the order of several centimeters (Reid et al., 2019). While Global Navigation Satellite Systems (GNSS) can often meet this requirement, they are rendered unreliable in dense urban environments due to signal occlusion and severe multi-path effects. However, these same environments offer a high density of stable, structured features—such as building facades, road surfaces, road signs—that can serve as geospatial landmarks for localization, providing a viable alternative to GNSS.

The dominant approach for achieving the required centimeter-level accuracy, without GNSS, relies on matching vehicle sensor data to a pre-built digital map, with High-Definition (HD) maps emerging as the state-of-the-art solution (Ghallabi et al., 2019, Chen et al., 2021). These multi-layered representations typically contain a rich set of geometric and semantic features, including dense 3D point clouds, road networks, poles, and curbs. HD maps enable a variety of localization methods, including Light Detection and Ranging (LiDAR)-based point-voxel (Shi et al., 2020) or segment matching (Dubé et al., 2017), as well as vision-based techniques that utilize features like road markings (Zhao et al., 2024) or traffic lights (Zhu et al., 2025). Multi-modal approaches fusing camera and LiDAR data also see common use with HD maps (Zuo et al., 2020).

Despite their proven efficacy, the creation and continuous maintenance of HD maps at scale incur prohibitive costs, as the map must be frequently updated to reflect changes in the real world (Lee and Ryu, 2024). This high overhead has inspired increasing research into localization using low-detail, abstract maps that are widely available and simpler to maintain. Examples include OpenStreetMap (OSM) (Elhousni et al., 2022) and maps derived from building information, such as 2D building footprints (Javanmardi et al., 2017) or their 3D extrusions (Vogel et al., 2018).

A core challenge in this domain is bridging the gap between dense, noisy, and dynamic sensor measurements and the sparse, abstract nature of the reference map. For 2D maps, common methods involve generating a simulated sensor view via raycasting and matching it against the real scanner data (Javanmardi et al., 2017, Elhousni et al., 2022); others match LiDAR data directly to satellite or aerial imagery (Lee and Ryu, 2024). While promising, these methods do not fully exploit the 3D geometry of the scene, limiting their ability to utilize full spatial information and to perform localization in 3D.

To address this, (Vogel et al., 2018, Moftizadeh, 2024) have utilized publicly available LoD2 building models and Digital Terrain Models (DTM) to benefit from abstract 3D reference maps for LiDAR-matching. However, recent work has highlighted feature extraction quality from raw sensor data as a key limitation, especially for achieving accurate and robust pose estimation (Wahbah et al., 2025).

This paper aims to address this gap by introducing a novel feature extraction pipeline that fuses geometric LiDAR data with semantic information from co-registered cameras. Our method projects semantic labels from an efficient 2D segmentation network (DDRNet-23) onto the 3D point cloud, which allows the robust and sequential extraction of the map's features: the ground plane (DTM), and building facades (LoD2).

Furthermore, the quantitative evaluation of feature extraction in abstract maps remains critical due to the absence of ground-truth (GT) solutions. Standard ground-truth generation methodologies, which rely on manual annotation (e.g., nuScenes

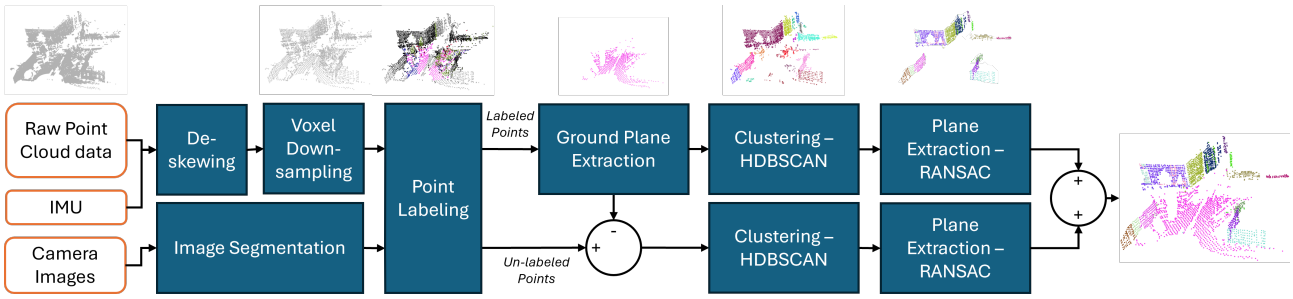


Figure 1. Overview of the proposed semantic-geometric feature extraction pipeline.

(Caesar et al., 2020), Waymo (Sun et al., 2020)), are unsuitable for this task. The abstract nature of the map features prevents a direct, one-to-one correspondence with the raw sensor data, rendering manual labeling of ”ground-truth features” infeasible for a human annotator. This absence of a benchmark makes it difficult to rigorously evaluate feature extraction performance in abstract maps.

We overcome this challenge with our second contribution: a novel framework for generating a benchmark of GT feature labels. We employ a multi-stage Human-in-the-Loop (HITL) process to establish a statistically robust 2D coarse pose for each sensor frame. This involves a two-stage ”nano-task” system where annotators identify correspondences between the point cloud and the reference map in a Bird’s-Eye View (BEV). This coarse pose is subsequently refined into a full 3D ”ground-truth pose” via point-to-plane ICP, and then validated using a final human inspection step. To the best of our knowledge, this represents the first quantitative evaluation of feature extraction with abstract LoD2/DTM maps.

This paper is structured as follows: Section 2 details the proposed semantic-geometric feature extraction pipeline. Section 3 describes the complete multi-stage ground-truth generation framework. Section 4 presents our experimental setup, validation of the generated ground-truth, and the performance results of our feature extraction pipeline. Section 5 discusses the results, and outlines system limitations. Section 6 concludes the paper with a summary of our contributions and an outlook on the future research direction.

2. Feature Extraction Pipeline

This section details our first contribution: a pipeline that processes raw dense sensor data from a dynamic scene into a set of planar features. The pipeline first robustly extracts the primary ground plane using semantically labeled points, thus simplifying the scene. Then a semantic filter discards all points corresponding to object classes not present in the map’s abstract representation (e.g., vehicles, pedestrians, vegetation). Subsequently, a geometric filtering stage employs clustering (HDBSCAN) and robust plane fitting (RANSAC) on the remaining points to isolate the large-scale vertical planes that directly correspond to the LoD2 building facades. This hybrid semantic-geometric approach yields a high degree of robustness to segmentation errors and measurement outliers. An overview of the system steps is shown in Figure 1.

2.1 Pre-processing and Semantic Fusion

The pipeline starts by ingesting camera images and LiDAR point clouds. The images are processed by an efficient semantic segmentation network. The proposed pipeline does not require a highly precise, pixel-accurate segmentation mask. This relaxed constraint is feasible because our feature extraction focuses on large-scale structures, while minor inaccuracies from the perception module are effectively mitigated by subsequent geometric processing steps. For our pipeline, we utilize the weights from the official DDRNet-23 (Pan et al., 2023) release, pretrained on the Cityscapes dataset (Cordts et al., 2016), to segment the road-scene images into 19 classes. We chose this network due to its high Intersection over Union (IoU) score, while achieving real-time capabilities (Lei et al., 2025).

Concurrently, the raw LiDAR point cloud, P_{raw}, is first undistorted to correct for motion-induced skew. This is achieved by continuously integrating high-frequency Inertial Measurement Unit (IMU) measurements to track the sensor’s ego-motion during a single scan acquisition, following the kinematic model detailed in (Xu and Zhang, 2021). By applying linear interpolation to estimate the pose of the sensor at the per-point timestamps, we project the raw points into a unified, motion-compensated local frame. The resulting geometrically consistent point cloud is then downsampled using a coarse voxel grid. This filtering step serves a dual purpose: it provides computational advantages by reducing point density, and it functions as a spatial low-pass filter. This effectively isolates the dominant structural components of the environment by averaging out sensor noise and suppressing minor architectural details (e.g., window ledges, water pipes, signs). Consequently, this focuses the data on the large-scale geometries that best align with the structures represented in the target abstract features.

Utilizing the preexisting extrinsic calibration and the overlapping FoV between the camera and LiDAR, the 2D segmentation mask is then projected onto the 3D voxelized point cloud (Vora et al., 2020), ”painting” all points within the shared field of view (FoV) with semantic labels. This divides the cloud into a set of labeled points, $P_{labeled}$, and unlabeled points, $P_{unlabeled}$ (those outside the shared FoV).

2.2 Geometric Feature Extraction

The extraction is a sequential process that first removes the ground-plane and subsequently extract building facades.

2.2.1 Ground Plane Removal To extract the DTM feature, a RANSAC plane-fitting algorithm is applied to points in $P_{labeled}$ marked as ”ground” or ”street.” This process generates a per-frame, single-plane model; this is a valid assumption within the context of urban canyons, where local street segments remain predominantly planar. This semantic guidance yields a highly robust ground plane model, π_{ground}. This plane is then extended to the entire point cloud, and all inlier points (from both $P_{labeled}$ and $P_{unlabeled}$) are removed, simplifying the scene.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-25-2026

2.2.2 Semantic Filtering From the remaining non-ground points, a critical semantic filter is applied. All points in $P_{labeled}$ belonging to classes that do not exist in the abstract map, such as ‘vehicles‘, ‘humans‘, ‘riders‘, ‘vegetation‘, and ‘fences‘, are discarded. This step leaves only points that are labeled as ‘buildings‘ and are highly likely to be part of the target features.

2.2.3 LoD2 Planes Extraction The remaining feature points, $P_{features}$, consist of the semantically filtered labeled points plus all unlabeled non-ground points. These are processed to find LoD2 facades:

1. **Clustering:** To enable efficient downstream processing, we cluster the points into distinct groups, thereby reducing the overall computational time. We employ the HDBSCAN algorithm (McInnes et al., 2017) for this task. We set the minimum cluster size C_n to be large enough to filter out smaller objects like trucks or minor walls that may have remained after the semantic filter. Additionally, we set the C_ϵ such that the algorithm is able to connect segments and form coherent clusters despite gaps caused by occlusion.
2. **Plane Fitting:** A RANSAC algorithm is applied on each cluster to extract the planar surfaces. A verticality constraint is applied, where only planes deviating less than $\epsilon_\perp$ from the vertical axis to the ground are preserved. The RANSAC distance threshold τ_{dist} is chosen to reject separate fixtures (e.g., balconies, window sills) while being large enough to include points on a single, slightly non-planar facade (e.g., decorative bricks).

The final output is a set of features:

$$\mathcal{F} = \{\pi_{ground}, \pi_{facade_1}, \pi_{facade_i}, ..., \pi_{facade_N}\} \quad (1)$$

which correspond directly to the DTM and LoD2 map primitives.

3. Ground-Truth Generation

Owing to the abstract nature of the map features, a novel strategy for GT benchmark generation is required for the quantitative evaluation of the feature extraction pipeline. The following section outlines our HITL approach for this purpose. First, we establish a robust 6-DoF ”GT pose” for each measurement epoch, and then we use this pose to automatically generate the GT feature labels. The process is divided into four primary stages: (1) HITL coarse 2D pose alignment, (2) 3D GT pose refinement, (3) GT pose validation, and (4) final GT feature label generation.

3.1 Stage 1: HITL Coarse 2D Pose Alignment

The first stage generates a statistically robust 2D coarse pose (x, y, θ) in the map coordinate system using a HITL process. While the INS pose cannot be strictly relied on in urban canyons to achieve our centimeter-level global accuracy, it is sufficient to establish a prior correspondence for the human annotators to enhance.

3.1.1 Nano-Task Annotation Annotators use a 2D Bird’s-Eye View (BEV) tool displaying both the LiDAR point cloud and the LoD2 building polygons as registered by the prior. The process is divided into two nano-tasks:

- **Task 1:** Annotators collaboratively label potential features (e.g., building corners) in the raw point cloud, P_{lidar}. This yields a set of source points $\{s_1, \dots, s_N\}$ where $s_i \in \mathbb{R}^2$.
- **Task 2:** For each source point s_i, multiple annotators independently and repeatedly identify the corresponding corner point on the LoD2 map polygons. This produces a set of R target points $\{t_{i1}, \dots, t_{iR}\}$ for each s_i. In addition, annotators can explicitly mark a correspondence as ”unsolvable” if no target point was found.

3.1.2 Statistical Aggregation on $SE(2)$ The human-derived annotations can be ambiguous and conflicting; thus, simply averaging them would yield a statistically invalid result. Instead, we generate a robust consensus using a non-parametric bootstrap and aggregation method that correctly accounts for the geometry of the transformation space.

First, we generate B bootstrap samples. For each sample $b \in \{1, \dots, B\}$, we resample with replacement from the target points $\{t_{i1}, \dots, t_{iR}\}$ for each source point s_i. We then aggregate these resampled targets using a robust, cluster-based mean. This method first finds a consensus cluster by identifying all points within a $10cm$ neighborhood of each other. If this cluster contains a majority of the resampled annotations, its mean is computed to create a single correspondence set $(s_i, \hat{t}_i^{(b)})$ for this bootstrap sample. Then, a RANSAC-based estimation is run on each set to compute a plausible 2D transformation $T_b \in SE(2)$. This results in a distribution of B transformations $\{T_1, ..., T_B\}$.

A direct arithmetic averaging of these transformations is not valid, as SE(2) lies on a non-Euclidean manifold. We therefore compute the approximate geometric median of this distribution by adapting a Weiszfeld-type algorithm (Weiszfeld, 1937) to operate on the Lie group. We initialize an estimate $\hat{T}^{(0)}$ (a random T_b) and iterate:

1. Compute the residual for each transformation in the tangent space at the current estimate $\hat{T}^{(k)}$, using the logarithm map:

$$\Delta_b = \log\left((\hat{T}^{(k)})^{-1} T_b\right) \in SE(2) \cong \mathbb{R}^3 \quad (2)$$

2. Compute weights w_b inversely proportional to the norm of the residuals:

$$w_b = \frac{1}{\max(||\Delta_b||_2, \epsilon)} \quad (3)$$

where ϵ is a small positive constant used for numerical stability.

3. Calculate the weighted mean of the residuals in the Lie algebra:

$$\overline{\Delta} = \frac{\sum_{b=1}^{B} w_b \Delta_b}{\sum_{b=1}^{B} w_b} \quad (4)$$

4. Update the estimate by mapping the mean residual back to the manifold via the exponential map:

$$\hat{T}^{(k+1)} = \hat{T}^{(k)} \exp(\overline{\Delta}) \quad (5)$$

This loop continues until convergence ($||\overline{\Delta}||_2 < \tau$). The final $\hat{T}$ is a robust 2D pose that represents the statistical consensus of all human annotations, and its associated covariance matrix $\Sigma = \frac{1}{B} \sum \Delta_b \Delta_b^\top$ quantifies the annotation uncertainty.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-25-2026 | © Author(s) 2026. CC BY 4.0 License.

3.2 Stage 2: 3D GT Pose Refinement

The robust 2D pose $\hat{T}$ from Stage 1, which consists of a 2D translation (t_x, t_y) and a rotation θ around the z-axis, is used to provide a high-confidence initial alignment for a full 3D refinement. This $SE(2)$ transformation is first embedded into an $SE(3)$ transformation matrix, T_{init}, by constraining the 6-DoF pose to the $x-y$ plane.

$$\hat{T} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & t_x \\ \sin(\theta) & \cos(\theta) & t_y \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow$$
$$T_{init} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 & t_x \\ \sin(\theta) & \cos(\theta) & 0 & t_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (6)$$

This 4x4 matrix, which assumes no rotation around the x and y axes, and zero z-translation, is applied to the raw point cloud. The feature extraction pipeline (described in Section 2) is then run on this coarsely aligned cloud to extract the planar ground and building features.

A point-to-plane Iterative Closest Point (ICP) algorithm is then used to match the extracted features from the point cloud to the corresponding DTM and LoD2 map features. The resulting 6-DoF transformation is designated as the GT Pose for that epoch.

3.3 Stage 3: GT Pose Validation

This stage is a critical step to ensure the quality of the final GT pose, mitigating any potential bias from using our own pipeline in the previous refinement stage. Every pose must pass two checks:

1. **Quantitative Check:** The final fitness score of the ICP algorithm from Stage 2 is stored. This score is defined as the ratio of the number of inlier points (points in the point cloud with map correspondences) to the total number of points in the point cloud, when transformed by that pose. The higher the fitness score, the better the alignment achieved. If the fitness score is below a predefined threshold, the pose is rejected as a low-quality match. This threshold is set to automatically filter out any low-quality matches, therefore, the threshold can be chosen from a relatively wide range. Provided the threshold is not restrictive enough to reject valid poses, its primary effect is on the volume of data passed to the qualitative check. In our case, a value of 0.75 was chosen.

2. **Qualitative Check:** This final human-in-the-loop inspection is essential because the quantitative check alone is insufficient. A high pose fitness score can confirm a geometric fit, but it cannot guarantee that the pose converged to the true global location (e.g., it might align perfectly with an adjacent, geometrically similar building). Therefore, a final human visual inspection is performed as the ultimate check to confirm that the point cloud, when transformed by the GT pose, is unambiguously aligned with the correct map features.

Only poses that pass both checks are used in the final benchmark.

3.4 Stage 4: GT Feature Label Generation

In the final stage, the validated GT poses are processed to automatically label the benchmark dataset. The GT poses are applied to their corresponding raw point clouds, P_{raw}, to align them with the LoD2/DTM map, creating P_{gt}. For each point $p \in P_{gt}$, we calculate its distance to the nearest map feature.

- If the point is within a 0.1-meter tolerance of an LoD2 building surface, it is labeled as **”building”**.
- If the point is within a 0.1-meter tolerance of the DTM ground surface, it is labeled as **”ground”**.
- All other points are labeled as **”clutter”**.

A $0.1m$ tolerance is applied to mitigate LiDAR and undistortion artifacts while discarding points unrelated to map features.

4. Experiments and Results

Our experimental setup consisted of a vehicle equipped with two Livox HAP LiDARs, mounted on the top of the vehicle and oriented at $\pm 40^{\circ}$ from the direction of travel to create a wide, non-overlapping FoVs. Each LiDAR was co-registered with a Basler a2A2840-67g5cBAS camera aimed in the same direction. An OxTS AV200 INS with dual-antenna GNSS provided the initial (”rough”) pose estimates, which were used as the starting point for the HITL annotation process.

The dataset was collected during a drive of multiple concentric loops through dense urban canyons in Hannover, Germany. The abstract reference map, consisting of LoD2 building models and a DTM, was sourced from the (Landesamt für Geoinformation und Landesvermessung Niedersachsen (LGLN), 2024) open data Portal. Our final evaluation dataset consists of 297 randomly chosen epochs, each with corresponding sensor data and an initial GNSS pose.

The performance of the feature extraction pipeline is governed by several key parameters, summarized in Table 1. These values were deliberately chosen based on our setup to match the pipeline’s objective: to ignore fine-grained sensor details and robustly extract the large-scale, abstract features present in the LoD2/DTM map, rather than being over-tuned to a specific dataset.

Table 1. Key parameters of the feature extraction pipeline.

Parameter	Value	Reasoning
Voxel Size	1.0 m	Filters fine-grained details (e.g., ledges) to focus on large-scale map structures.
HDBSCAN C_n	25 pts	Based on voxel size, ensures a minimum detectable surface of $\approx 25\,\mathrm{m}^2$, filtering out planar, non-building objects (e.g., trucks).
HDBSCAN C_ϵ	2.0 m	Allows point clusters to form across gaps caused by sensor occlusions.
RANSAC τ_{dist}	0.25 m	Rejects fixtures (e.g., balconies) while including points on a single, non-planar facade.

4.1 Evaluation Metrics

To evaluate our feature extraction pipeline, we compare the set of detected map features against the set of ground-truth (GT) map features on a per-frame basis. We define the ground-truth features set (G) as the set of map surfaces visible in the benchmark, and the detected features set (D) as the set of map surfaces identified by our pipeline. We then compute the standard metrics:

$$\text{Precision} = \frac{|G \cap D|}{|D|} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{|G \cap D|}{|G|} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{IoU} = \frac{|G \cap D|}{|G \cup D|} \tag{9}$$

Here, Precision quantifies the reliability of our detections (low False Positives, FP), while Recall measures the completeness of our pipeline (low False Negatives, FN). Finally, IoU provides a comprehensive metric for accuracy by measuring the ratio of overlap between predicted and ground-truth point sets, accounting for both the FP and FN.

A critical challenge in this evaluation is the discrepancy in point density. The GT feature labels are generated by labeling the original, raw point cloud, which is dense. However, our feature extraction pipeline applies a coarse 1.0 meter voxel filter, which significantly reduces point density. To ensure a fair and valid comparison, we first normalize the benchmark. Before applying any metrics, the raw GT point cloud (with its feature labels) is processed through the exact same 1.0-meter voxel filter used in our pipeline.

Finally, we should distinguish between significant features and noise. We define a minimum threshold for a valid feature in both the G and D sets. A map surface is only included in G or D sets if it is associated with at least 25 points. This value is chosen to directly match the C_n parameter from our pipeline's HDBSCAN step, as this represents the minimum significant feature our pipeline is designed to detect.

4.2 Ground-Truth Validation Results

The 3D GT pose alignment statistics for the 297 epochs are detailed in Table 2. The mean ICP fitness score was 0.9346, indicating that, on average, over 93% of the extracted feature points from the GT-poses scans align with the abstract map. The mean inlier RMSE was $0.3409m$ with a low standard deviation of $2.84cm$, confirming the high precision of the poses. The distributions of these metrics (Figure 2) show a strong positive skew in fitness scores and a tight, Gaussian-like distribution in RMSE, confirming the reliability of the GT baseline.

It should be noted that the alignment fitness score cannot reach the theoretical 100%, even with ideal measurements. This is because the LoD2 map features are generalized representations, while the LiDAR scans capture the full geometric complexity of the real-world environment.

4.3 Feature Extraction Performance

We evaluated our feature extraction pipeline against the validated 297-frame benchmark. The overall performance, summarized in Table 3, demonstrate the high accuracy and stability of our proposed pipeline. We report metrics in three ways:

Table 2. Alignment statistics for the validated GT poses.

Metric	Fitness Score	Inlier RMSE (m)
Mean	0.9346	0.3409
Std. Dev	0.0417	0.0284
Median	0.9488	0.3356
Min	0.7681	0.2923
Max	0.9926	0.4335

”Overall” (calculated from the total sum of TPs, FPs, and FNs), ”Mean Per-frame”, and ”Median Per-frame”.

The Median Per-frame Precision is 100.00%, which indicates that in at least 50% of all evaluated frames, the pipeline produced zero, feature-wise, false positives. This demonstrates an extremely high level of reliability in the typical case. The ”Mean Per-frame” precision of 95.14% is only slightly lower, confirming that a small subset of frames with known failure modes are responsible for nearly all false positives.

Furthermore, the pipeline's recall is very stable: the ”Median Per-frame Recall” (83.33%), ”Mean Per-frame” recall (82.37%), and ”Overall” recall (83.74%) are all closely clustered. The negligible divergence between the mean and median recall suggests that performance is highly consistent across the dataset, not skewed by a few low-performing outlier frames.

Table 3. Overall and per-frame feature extraction performance.

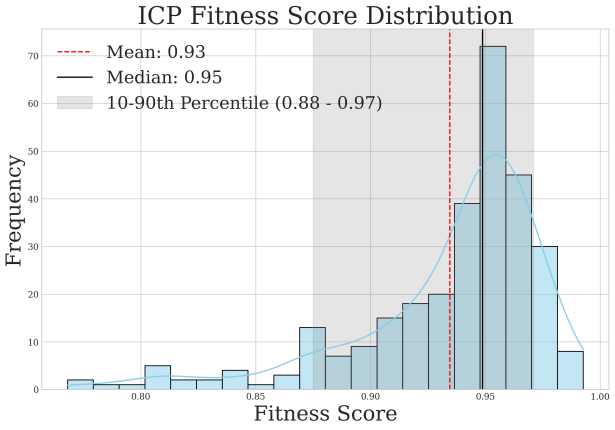
Metric	Value
Total GT Features (in 297 valid packets)	4232
Total Detected Features	3729
Overall Geometry Precision	95.04%
Overall Geometry Recall	83.74%
Overall Geometry IoU	80.24%
Mean Per-frame Precision	95.14%
Mean Per-frame Recall	82.37%
Mean Per-frame IoU	79.41%
Median Per-frame Precision	100.00%
Median Per-frame Recall	83.33%
Median Per-frame IoU	80.00%

The analysis in Figure 3 provides deeper insight. Figure 3a shows a near-perfect overlap between the ground-truth and detected histograms. The mean (114.8 vs 113.8) and median (50 vs 50) point counts are almost identical, proving that our pipeline is statistically unbiased and detects features across all scales, not just large, simple ones.

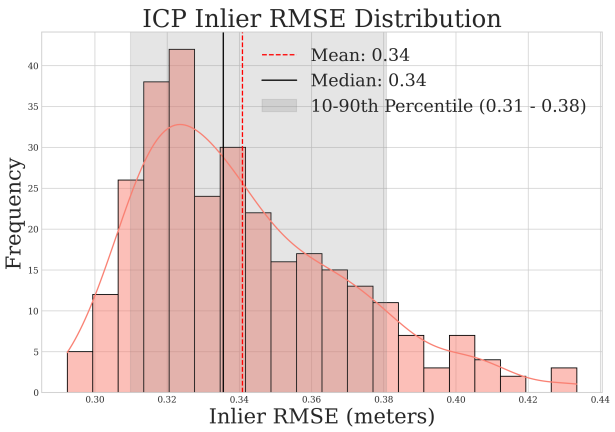
On the other hand, Figure 3b provides a validation for our selection of C_n parameter. The analysis reveals that the pipeline's recall is highly correlated with the statistical significance of the GT features. Recall is low for features defined by fewer than 20 points, demonstrating our pipeline's ability to intentionally and correctly reject small, ambiguous, or noisy surfaces that are unreliable for localization, while maintaining high recall of 83.7% precisely at our 25-point threshold, confirming this parameter is well-tuned to balance sensitivity and robustness. Critically, for all large features (defined by > 80 points), the pipeline achieves 100% recall, proving it reliably captures the most significant geometric structures that should most benefit the downstream geo-referencing process.

4.4 Qualitative Assessment

To complement the quantitative metrics, we provide a qualitative assessment of the feature extraction pipeline. As illustrated

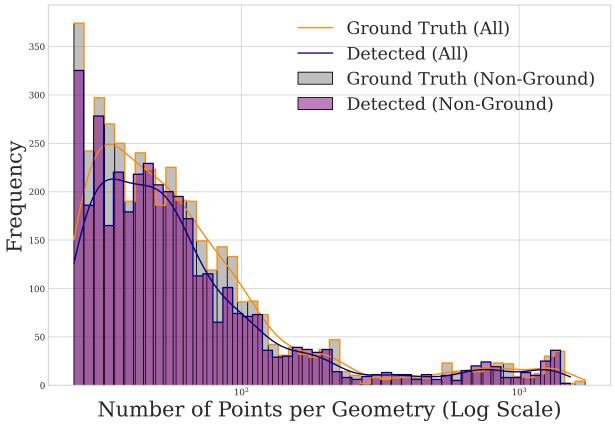


(a) Distribution of ICP Fitness Scores

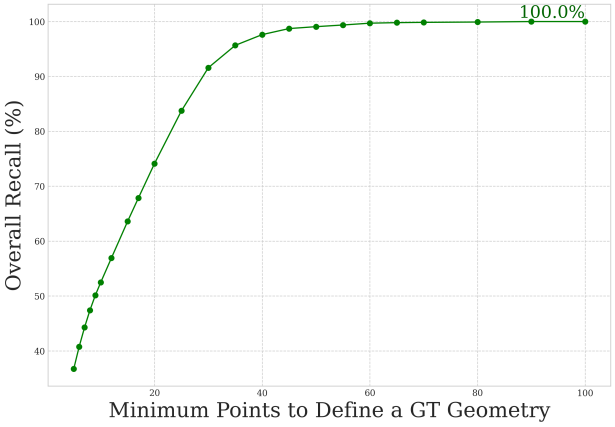


(b) Distribution of ICP Inlier RMSE

Figure 2. Distributions of ICP fitness scores 2a and inlier RMSE 2b for the 297 validated epochs in the GT benchmark. The red dashed line indicates the mean.



(a) Distribution of GT and detected features



(b) Recall as a function of GT minimum detected points

Figure 3. Feature extraction performance analysis. Figure 3a provides a comparison of the size distributions for ground-truth and detected features, showing a strong positive correlation. Figure 3b shows the recall as a function of the minimum number of points required to define a valid ground-truth feature.

in Figure 4, the proposed method effectively isolates the LoD2 and DTM geometric structures from dense, raw point clouds within complex urban environments.

The figure demonstrates the robustness of our multi-stage approach: the isolation of the ground plane (Stage 2), the high-fidelity clustering of candidate points via HDBSCAN (Stage 3), and the final extraction of planar features with RANSAC (Stage 4). Across diverse samples (A–E), the pipeline consistently maintains geometric integrity despite varying point densities and surface occlusions.

Consistent with our quantitative findings, minor FN and FP discrepancies are occasionally endured. For instance, in specific cases such as Stage 3 and 4 of Sample D, certain planar points were misclassified as planar due to the combined effects of occlusion and the downsampling resolution. However, these instances represent isolated outliers that do not significantly impact the overall efficacy of the developed pipeline.

5. Discussion

The experimental results in Section 4 provide strong quantitative validation for our two primary contributions. The following discussion addresses the paper’s central methodological challenges, and the system’s viability for real-world application.

5.1 Addressing Methodological Circularity

A primary concern for this work is the methodological circularity, we use our feature extraction pipeline within the GT generation framework to create a benchmark, which is then used to evaluate the performance of our feature extraction pipeline.

We argue that this circularity is broken by the approach in Section 3.3:

1. Our pipeline was used in Stage 2 only as an intermediate tool to provide a robust initialization for a standard ICP algorithm.
2. The final GT Pose is the output of the ICP, not the output of our pipeline.
3. This pose was then subjected to an independent quality check (Stage 3) and validated against the map by a human annotator.
4. The generated ground-truth geometries are extracted from the raw point cloud, based on their alignment with the map features, independent of our pipeline.

Additionally, the results of the high mean ICP fitness of 0.9346, and mean ICP RMSE of $0.3409\ m$ are the independent proof

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-25-2026

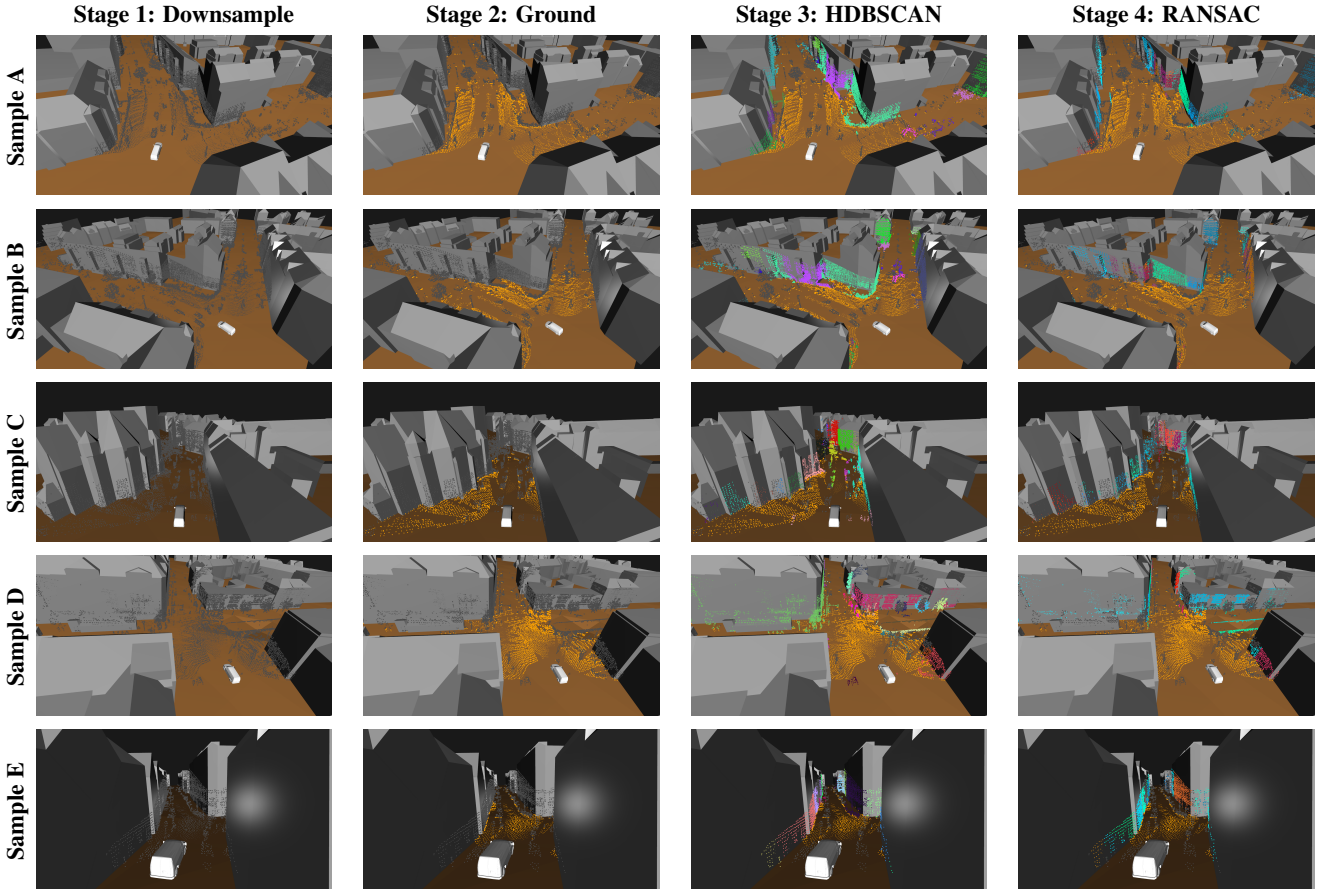


Figure 4. Qualitative assessment of the feature extraction pipeline across five distinct samples (A–E). Each row demonstrates the sequential processing stages visualized along the LoD2 and DTM layers: (1) downsampled point cloud (gray); (2) extracted ground points (gold); (3) HDBSCAN-based point clustering (color-coded); and (4) final RANSAC extracted geometric planes (color-coded). The vehicle model is included for orientation purposes only; it is not to scale and does not represent an exact physical pose

that the final GT poses are of high quality, regardless of the tool used to initialize the ICP.

Furthermore, if the initial pipeline fails to extract sufficient features during Stage 2, the ICP alignment fails to converge; consequently, the epoch is flagged and discarded by the validation threshold in Stage 3. Such failures typically occur in feature-sparse environments, such as open parking lots, or where the abstract map lacks corresponding elements, like tunnels and bridges. While data from featureless areas can be omitted as it doesn’t represent our abstract map, incorporating structural models for tunnels and bridges into the map priors represents a compelling opportunity for future work.

5.2 Viability for Real Time Applications

While the focus of this paper is on the robustness of feature extraction and the benchmark generation, the pipeline was designed with its real-time suitability in mind. The semantic segmentation network, DDRNet-23, was chosen for its high efficiency and is reported to run at rates higher than 50 FPS (approx. 20 ms) (Lei et al., 2025). The remainder of the feature extraction workflow (voxelization, ground removal, filtering, and clustering) was measured to execute in approximately 80 ms using a CPU-based non-optimized Python implementation on a ‘i5-12500H‘ processor.

The total processing time of approximately 100 ms, which matches the 10 Hz update rate of common LiDARs, is insufficient for real-time localization applications due to the resulting latency and the absence of a safety margin (European Commission, 2022). Nevertheless, the geometric processing currently runs on a non-optimized Python prototype and generic hardware. Translating this prototype to a C++ implementation utilizing parallelization (e.g., CUDA or multi-threading for the geometric clustering) is the intended path to meet the ∼20-30 ms real-time safety margins. This would allow the pipeline to meet, and likely exceed, standard real-time requirements. Additionally, the standard segmentation network can be replaced with a fine-tuned network for our specific task, further reducing the processing time.

5.3 Parameter Selection and System Limitations

Despite the high performance, the framework has limitations. The first is parameter generalizability. The key parameters, such as Voxel Size, C_n, and τ_{dist} were tuned for our specific sensor suite (dense, side-mounted LiDARs) and test environment. These would likely require re-tuning for different sensor modalities (e.g., sparser, 360° LiDARs) or environments. For instance, when utilizing a sparser LiDAR configuration, the overall point density on environmental surfaces is significantly lower. To adapt, the HDBSCAN clustering minimum cluster C_n size should be decreased to ensure that the now-sparser building facades are not inadvertently filtered out, while the cluster distance threshold C_ϵ may need to be increased to effectively bridge the larger spatial gaps between individual points.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-25-2026 | © Author(s) 2026. CC BY 4.0 License.

We observed that the Voxel Size had the largest effect on performance; however, the system produced comparable results for a wide range of values (0.25 m to 2.0 m), provided the other parameters were scaled in accordance with the pipeline's core logic of filtering for large-scale features. When dealing with sparse sensor configurations, adjusting this voxel size requires a careful balance to prevent over-filtering the already limited geometric data, and parameters like the RANSAC distance threshold τ_{dist} must be proportionately scaled to capture slightly non-planar facades reliably. Nonetheless, a formal sensitivity analysis across different sensors and environments is required.

Finally, due to the unified ground modeling approach, the system might be vulnerable to complex environments that contain steep topological changes; however, these situations are rare in dense urban environments, which are the main focus in our study.

6. Conclusion and Future Work

This paper presented two core contributions to address the challenge of robust feature extraction in abstract LoD2/DTM maps. First, we introduced a novel semantic-geometric feature extraction pipeline that robustly identifies map-correlative features from dense, dynamic sensor data. Second, we developed a multi-stage HITL framework that leverages human contextual understanding alongside automated labeling to generate a high-quality benchmark dataset.

Our experiments validate both contributions. The GT framework produced a 297-frame benchmark of high-quality poses, verified by a mean ICP fitness of 0.9346 and a mean RMSE of 0.341 m. Evaluated against this benchmark, our feature pipeline demonstrated high performance, achieving 95.14% precision and 83.74% recall while remaining statistically unbiased across all feature scales.

For future work, the most direct next step is to expand our GT benchmark to encompass the majority of our dataset, and to integrate the validated feature extraction pipeline into a full real-time localization framework (e.g., a Kalman filter or a factor-graph optimizer). While the current study deliberately focuses on validating the isolated feature extraction performance, evaluating the downstream, end-to-end vehicle pose estimation accuracy using these features is the immediate next step in our research. Furthermore, we plan to address the system's current limitations by incorporating temporal information. This will enable the selected filter to better distinguish between static map features and static false positives (e.g., temporary scaffolding or a parked truck), thereby improving the overall robustness of the system.

Future work will focus on expanding the benchmark to the full extent of our collected data and integrating the validated feature extraction pipeline into a real-time localization framework, such as a factor-graph optimizer. While this study focused on isolating and validating feature extraction performance, evaluating downstream end-to-end vehicle pose estimation is necessary for valorizing LoD2 and DTM in autonomous driving applications. Furthermore, we plan to incorporate temporal information to improve robustness, allowing the system to better distinguish between static map features and persistent non-map structures, such as temporary scaffolding or parked vehicles.

Acknowledgment: This work was supported by the AutoMap project, which is funded by the German Federal Ministry of Transport (BMV).

References

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuScenes: A Multimodal Dataset for Autonomous Driving . *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos, CA, USA, 11618–11628.

Chen, G., Lu, F., Li, Z., Liu, Y., Dong, J., Zhao, J., Yu, J., Knoll, A., 2021. Pole-Curb Fusion Based Robust and Efficient Autonomous Vehicle Localization System With Branch-and-Bound Global Optimization and Local Grid Map Method. *IEEE Transactions on Vehicular Technology*, 70(11), 11283-11294.

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

Dubé, R., Dugas, D., Stumm, E., Nieto, J., Siegwart, R., Cadena, C., 2017. SegMatch: Segment based place recognition in 3D point clouds. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 5266–5272.

Elhousni, M., Zhang, Z., Huang, X., 2022. LiDAR-OSM-Based Vehicle Localization in GPS-Denied Environments by Using Constrained Particle Filter. *Sensors*, 22(14). https://www.mdpi.com/1424-8220/22/14/5206.

European Commission, 2022. Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 Laying Down Rules for the Application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as Regards Uniform Procedures and Technical Specifications for the Type-Approval of the Automated Driving System (ADS) of Fully Automated Vehicles. *Official Journal of the European Union*, L 221, 1–64.

Ghallabi, F., El-Haj-Shhade, G., Mittet, M.-A., Nashashibi, F., 2019. LIDAR-based road signs detection for vehicle localization in an HD map. *2019 IEEE Intelligent Vehicles Symposium (IV)*, 1484–1490.

Javanmardi, E., Javanmardi, M., Gu, Y., Kamijo, S., 2017. Autonomous vehicle self-localization based on multilayer 2D vector map and multi-channel LiDAR. *2017 IEEE Intelligent Vehicles Symposium (IV)*, 437–442.

Landesamt für Geoinformation und Landesvermessung Niedersachsen (LGLN), 2024. Open geodata hub - wertermittlung. Webpage.

Lee, S., Ryu, J.-H., 2024. Autonomous Vehicle Localization Without Prior High-Definition Map. *IEEE Transactions on Robotics*, 40, 2888-2906.

Lei, X., Chen, Z., Yu, Z., Jiang, Z., 2025. BENet: Boundary-Enhanced Network for Real-Time Semantic Segmentation. *The Visual Computer*, 41(1), 229-241.

McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), 205. https://doi.org/10.21105/joss.00205.

Moftizadeh, R., 2024. *Advanced particle filtering for vehicle navigation based on collaborative information*. DGK: C (Dissertationen), Heft Nr. 941, München.

Pan, H., Hong, Y., Sun, W., Jia, Y., 2023. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes. *IEEE Transactions on Intelligent Transportation Systems*, 24(3), 3448-3460.

Reid, T. G. R., Houts, S. E., Cammarata, R., Mills, G., Agarwal, S., Vora, A., Pandey, G., 2019. Localization Requirements for Autonomous Vehicles. *SAE International Journal of Connected and Automated Vehicles*, 2(3), 173–190.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H., 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10526–10535.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D., 2020. Scalability in perception for autonomous driving: Waymo open dataset. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2443–2451.

Vogel, S., Alkhatib, H., Neumann, I., 2018. Iterated Extended Kalman Filter with Implicit Measurement Equation and Nonlinear Constraints for Information-Based Georeferencing. *2018 21st International Conference on Information Fusion (FUSION)*, 1209–1216.

Vora, S., Lang, A. H., Helou, B., Beijbom, O., 2020. Pointpainting: Sequential fusion for 3d object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4603–4611.

Wahbah, M., Ramme, L., Ernst, D., Vogel, S., Neumann, I., Alkhatib, H., 2025. Geo-referencing Autonomous Vehicles Using LoD2 and HD Maps: Performance Assessment in Simulated Urban Environments. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-G-2025, 915–922. https://isprs-annals.copernicus.org/articles/X-G-2025/915/2025/.

Weiszfeld, E., 1937. Sur le point pour lequel la Somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43, 355-386.

Xu, W., Zhang, F., 2021. FAST-LIO: A Fast, Robust LiDAR-inertial Odometry Package by Tightly-Coupled Iterated Kalman Filter. *IEEE Robotics and Automation Letters*, 6(2), 3317-3324.

Zhao, L., Liu, Z., Yin, Q., Yang, L., Guo, M., 2024. Towards robust visual localization using multi-view images and hd vector map. *2024 IEEE International Conference on Image Processing (ICIP)*, 814–820.

Zhu, F., Zhou, R., Chen, W., Yu, M., Zhang, X., 2025. Fusing Information From Multi-Sensors and High-Definition Maps for Continuous and Precise Positioning in Autonomous Driving Services. *IEEE Transactions on Intelligent Transportation Systems*, 1-19.

Zuo, X., Ye, W., Yang, Y., Zheng, R., Vidal-Calleja, T., Huang, G., Liu, Y., 2020. Multimodal Localization: Stereo Over LiDAR Map. *Journal of Field Robotics*, 37(6), 1003-1026.