Multi-Sensor based Perception and State Estimation for Autonomous Vehicles

Songming Chen

THÈSE DE DOCTORAT DE

L'ÉCOLE CENTRALE DE NANTES

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Automatique, productique et robotique*

Par
Songming CHEN

Multi-Sensor based Perception and State Estimation for Autonomous Vehicles

Thèse présentée et soutenue à l'Ecole Centrale de Nantes le 9 décembre 2022
Unité de recherche : UMR 6004, Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Paul CHECCHIN	Professeur des Universités, IUT Clermont Auvergne
Maan EL BADAOUI EL NAJJAR	Professeur des Universités, Université de Lille

Composition du Jury :

Présidente :	Marie BABEL	Professeure des Universités, INSA Rennes
Examinatrice :	Samia AINOUZ	Professeure des Universités, INSA Rouen Normandie
Dir. de thèse :	Vincent FRÉMONT	Professeur des Universités, École Centrale de Nantes

ACKNOWLEDGEMENT

This work was made possible through the financial support from the China Scholarship Council (CSC) and the scientific support from the École Centrale de Nantes (ECN) and Le Laboratoire des Sciences du Numérique de Nantes (LS2N).

First of all, I would like to express my gratitude to my thesis supervisor Prof. Vincent FRÉMONT for taking me on as a student, and giving me prompt encouragement and constructive advices throughout of the thesis. Without his mentorship in conducting research and writing papers, I would never have made such progress and successfully completed the dissertation. I would also like to acknowledge the thesis jury members, especially the thesis reviewers Prof. Paul CHECCHIN and Prof. Maan EL BADAOUI EL NAJJAR, for their time and effort devoted to reviewing my manuscript and giving the valuable feedbacks. My gratitudes also need to be expressed to the thesis examiners Prof. Marie BABEL and Prof. Samia AINOUZ, who travelled long-way and attended my thesis on-site. Their remarks and comments provoked me to have a good understanding of my thesis limitations, which may be improved in the future perspective work.

Then, I would like to thank the colleagues from the LS2N ARMEN team, who enriched my study and reserach life in École Centrale de Nantes. The regularly held group seminars provided the opportunities for us to communicate and learn from each other. I gained knowledge on practical programming tools from the presentation of Franco, Philipp, Shiyu and Andrea. While, the discussion with Minh-quan and Haixin on data-driven based approaches inspired me to leverage the semantic information for autonomous driving. The technical support from the research engineers David, Arnaud and Alexis ensured me to collect the data onboard and carry out the experiments without hassle.

Last but not least, I am deeply grateful to my family and friends for their continuous care, particularly during the outbreak of Covid-19. The love, patience and accompany of my girlfriend Xiaoxiao helped me to stay positive and confident through the hard moments of my thesis. And the encouragement and comfort of my mother Huiling, my father Yuguo, and all the rest of my family back in China paved the way for me to overcome the difficulties whenever met in my life.

Acronymes

ADAS	Advanced Driver-Assistance Systems
AV	Autonomous Vehicles
RANSAC	Random Sample Consensus
PTAM	Parallel Tracking and Mapping
ORB	Oriented FAST and Rotated BRIEF
SLAM	Simultaneous Localization And Mapping
LiDAR	Light Detection And Ranging
LOAM	Lidar Odometry and Mapping
SfM	Structure-from-Motion
SIFT	Scale-Invariant Feature Transform
SURF	Speeded Up Robust Features
BRIEF	Binary Robust Independent Elementary Features
BA	Bundle Adjustment
MOS	Moving Objects Segmentation
HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine
R-CNN	Regions with Convolutional Neural Networks
ROI	Regions of Interest
RPN	Region Proposal Network
YOLO	You Only Look Once
FVB	Flow Vector Bound
KLT	Kanade–Lucas–Tomas
FPS	Frames Per Second
DoF	Degree of Freedom
SAD	Sum of Absolute Differences
GNSS	Global Navigation Satellite Systems
ICP	Iterative Closest Point
GICP	Generalized-ICP

PFH	Point Feature Histograms
FPFH	Fast Point Feature Histograms
SHOT	Signature of Histograms of OrienTations
SAC-IA	Sample Consensus Initial Alignment
LSTM	Long Short-Term Memory
RMSE	Root Mean Square Error
RF	Relative Fitness
D-ICP	Direct Iterative Closest Point
F-ICP	Feature-based Iterative Closest Point
SD-ICP	Semi-Direct Iterative Closest Point
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FCN	Fully Connected Network
VFE	Voxel Feature Encoding
SSD	Single Shot Detector
BEV	Bird Eye View
FOV	Field Of View
DL	Deep Learning
IMU	Inertial Measurement Unit

List of Figures

List of Tables

Table of Contents

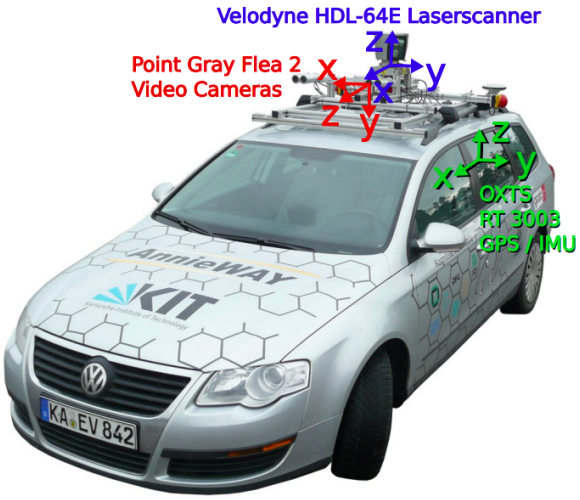
Chapter 1

Introduction

1.1 Research Background

In recent years, tremendous development has been made in the automotive industry. And substantial research efforts have been made to investigate the Advanced Driving Assistance System (ADAS) or Autonomous Vehicles (AV). Nowadays, the autonomous vehicles are usually equipped with several proprioceptive and exteroceptive sensors to perceive the surroundings, which promotes safety-critical self-driving. Specifically, the sensor measurements enable the autonomous vehicle to understand the scene context, on which basis, high-level tasks such as localization, obstacle avoidance and adaptive cruise control can be performed. Among the exteroceptive sensors, the cameras provide rich color and texture information of the scene. The visual features can be leveraged for 2D object detection, which can localize the objects on the image plane. Moreover, the features can be associated across the frames for relative pose estimation. In order to address the issue of depth loss, the multi-layer LiDAR can be used to acquire the range measurement. The LiDAR sensor is also invariant to the illumination conditions, which well complements the visual sensors. However, in complex urban environments, it is still challenging to have a comprehensive scene understanding. Due to the existence of moving objects, mutual occlusion and unexpected object movements tend to complicate the scene perception and ego-motion estimation tasks.

Driven by the advancement in hardware performance and corresponding software support, sensor fusion becomes a mainstream for the automotive perception system design. The sensor fusion-based architecture aims to leverage multiple sensors' measurements to empower the autonomous vehicle perception, so that a richer and more reliable representation of the surroundings can be obtained. Based on the accurate scene perception, the autonomous vehicle can accordingly implement the tasks such as self-localization, predictive path planning, etc. In this thesis, the automotive perception system which incorporates the cameras and LiDAR sensors is extensively studied (see Fig. 1.2). The



(a) The multi-sensor setup configured on the Volkswagen vehicle in the KITTI dataset

(b) The multi-sensor setup configured on the Renault vehicle in the LS2N dataset

Figure 1.1 – The exteroceptive sensors are attached onboard for experimental benchmarking purposes

exteroceptive Camera and LiDAR sensors have their own strengths and weaknesses under different working conditions. One main contribution of this thesis is to adaptively fuse the complementary vision-based and range-based sensors, where respective sensing modality estimation uncertainty is considered. Moreover, this thesis also bridges the gap between model-based and data-driven approaches, where the data-driven high-level semantic cues effectively complement the model-based ego-motion estimation in dynamic scenarios. Extensive experimental tests are carried out on both KITTI and LS2N datasets

(a) (b)

Figure 1.2 – Camera and LiDAR sensors are leveraged for scene perception and self-localization

(see Fig. 1.1). The proposed benchmarking scenarios include but are not restricted to parking, urban, and highway areas. The benchmarking results suggest the superiority of the multi-sensor fusion approach. Finally, the advantage of semantics involvement in ego-motion estimation and mapping processes is also highlighted.

1.2 Literature Review

The application of exteroceptive sensors to perceive the environment and estimate vehicle ego-motion can date back to the 1980s. In order to overcome the wheel slippage problem, the slider stereo-vision system is employed in [1] to estimate the 6 DoF ego-motion of the planetary rover. Due to the limited computation resources onboard, the terrain rover has to follow the stop-and-go motion to spare time for the image signal processing. With the 3D information of the surroundings, the planetary rover is capable of implementing the task of obstacle avoidance. This pioneering work [1] proposes a feature-based ego-motion estimation pipeline, which is classic and lays the foundation for modern applications. Afterwards, the vision-based perception system is equipped by the NASA Mars Exploration Rover [2] [3], which successfully ensures the planetary exploration safety and accelerates the rovers' navigation to the target area. However, the computation efficiency remains a bottleneck and the rover perception system can not be applied in real-time applications. The first real-time motion estimation architecture is proposed by Nister et al. in [4], where the five-point algorithm [5] is embedded in the Random Sample Consensus (RANSAC) [6] framework for efficient outlier rejection. Moreover, the 3D-2D reprojection error is for the first time used in [4] to estimate the camera relative motion, which successfully integrates the map points location and camera

poses into a single minimization function. To further improve the pose estimation accuracy and efficiency, the sliding-window based bundle adjustment approach is proposed in [7] by Mouragnon et al. The Jacobian matrix for the bundle adjustment problem renders a specific sparse structure, which ensures real-time implementation.

The modern vision-based perception system can not only estimate the camera poses but also capture the scene structure simultaneously. Among them, the Parallel Tracking and Mapping (PTAM) system [8] is a representative work, which initially divides the state estimation problem into front-end and back-end threads (see Fig. 1.3). Hereby, the state

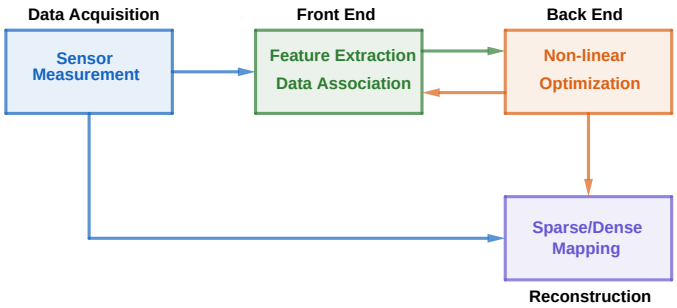


Figure 1.3 – Scheme of the front end and back end responsibilities for the state estimation problem

estimation refers to the vehicle poses and map points location estimation. On the one hand, the front-end thread is responsible for real-time feature tracking and relative pose estimation. On the other hand, the back-end thread performs the non-linear optimization to correct the map point positions and camera poses, which makes the state estimation process resilient to spurious data association. It needs to be noted that the back-end optimization does not require real-time implementation, and the thread synchronization can be performed when necessary, owing to the parallelization of the tracking and mapping process. Additionally, the keyframe-based pose estimation and mapping pipeline avoids data association and optimization for each frame, which saves substantial computing resources. It can be seen from Fig. 1.4 that the sparse map points are plotted to represent the scene. This map representation is quite compact and consists of the feature points (landmarks) observed from different view points, which is sufficient for the localization purpose. Nonetheless, the PTAM framework is just applicable in small scenes and there is no remedy when the feature tracking thread fails. Encouragingly, the vision-based perception system has made astonishing progress in the past decade, which enables autonomous

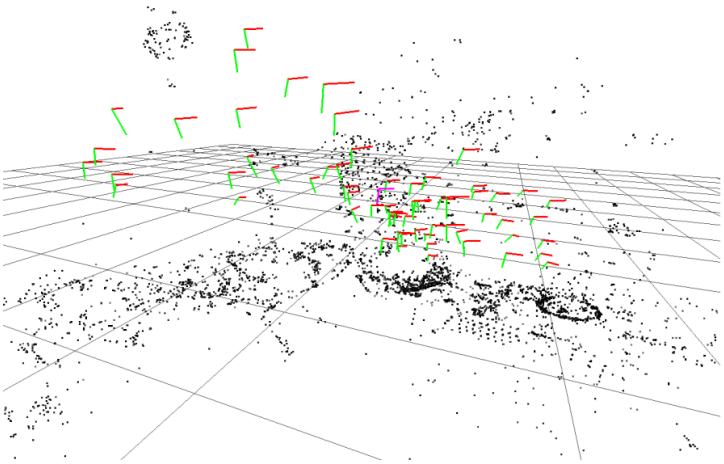


Figure 1.4 – The estimated keyframe poses and 3D map points with the PTAM framework [8]

vehicle large-scale exploration. ORB-SLAM [9] is one of the complete vision-based perception system, which is essentially an ORB [10] feature-based approach built upon on the PTAM framework. The ORB-SLAM pipeline consists of three main components, namely the feature tracking, local mapping and loop closing threads. The ORB feature is invariant to the scaling and rotation transformation, which ensures the robust feature tracking across frames. For severe viewpoint changes, the relocalisation module [11] will be activated to recover the feature tracking process. A bag of words place recognizer with ORB features is also proposed in [11] for loop closure detection, which will greatly reduce the accumulated drift. The ORB-SLAM2 system is then proposed in [12] to support the stereo and RGB-D cameras. With the stereo-vision system, the scale metric can be recovered. After the keyframe-based local and global optimization, the final trajectory tends to be accurate (see Fig. 1.5). Nevertheless, ORB-SLAM2 architecture relies on the sparse feature points for the scene mapping. This landmark-based scene representation only satisfies the localization purposes and is less appealing for the navigation tasks such as obstacle avoidance or path planning. For the navigation or path planning tasks, the dense map representation is more appropriate to avoid collision between two landmarks. According to the specific applications, the common dense map representation can be either 2D occupancy grid map [13] or 3D octomap [14], which are shown in Fig. 1.6 and Fig. 1.7. The dense maps can identify and present the drivable areas within the probabilistic framework, which favors the conservative and secure decision making.

Figure 1.5 – The estimated vehicle trajectory and reconstructed sparse map of the scene with ORB-SLAM2 [12]

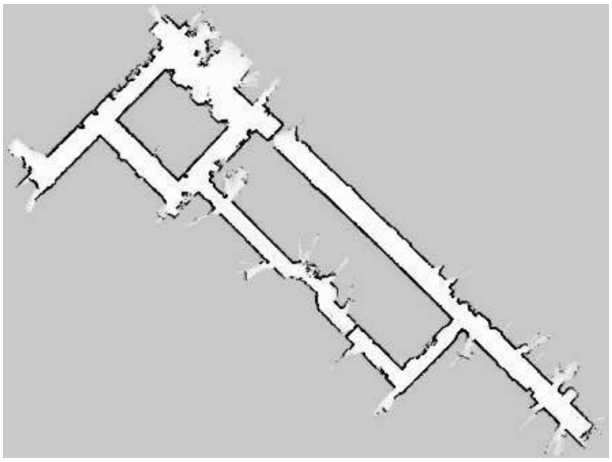


Figure 1.6 – Scheme of the 2D Occupancy Grid Map [13]

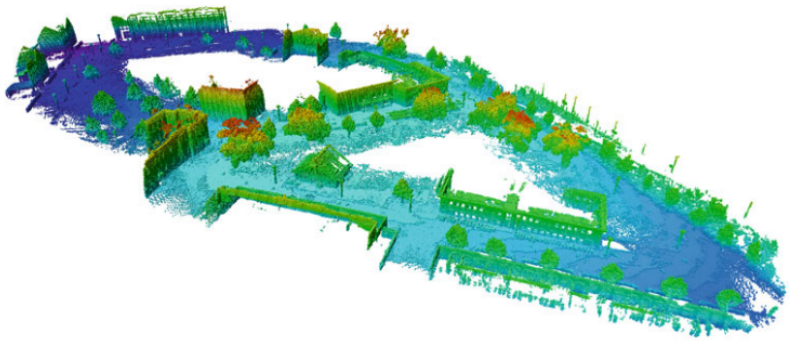


Figure 1.7 – Scheme of the 3D OctoMap [14]

Light Detection And Ranging (LiDAR) sensor is gaining the popularity in automotive industry, owing to its high precision and resolution for the scene perception. The LiDAR sensor measurements consist of dense 3D point clouds, which prepares the condition for dense scene reconstruction. Meanwhile, geometric features can be extracted from the dense point clouds for efficient relative transformation estimation. It is shown in LOAM [15] that the edge and planar features can be detected according to the curvature value. Then the point-to-edge and point-to-plane distances are minimized to calculate the scan-to-scan relative pose. In the back-end, the received LiDAR scans are matched with the local map to reduce the estimation drift (see Fig. 1.8). Furthermore, the real-time LiDAR-based loop closure is managed in the Cartographer framework [16], where the scan-to-submap matching constraints come from the branch-and-bound searching. The 2D Cartographer SLAM assumes a flat world, and the correlative scan matcher works well with the vertical direction implicitly defined. For 3D Cartographer SLAM, the inertial measurement unit is needed to measure the gravity.

The ego-motion estimation result reached by visual-SLAM algorithms can be enhanced via integrating LiDAR measurements. In [17], depth information from LiDAR measurements was utilized for visual feature tracking after LiDAR points being projected onto image frames. At the same time, visual semantic information was used for removing outliers and increasing the weights of static landmarks. Instead of using visual feature points, in [18] a SLAM system using visual photometric information was proposed. Its performance was enhanced with the involvement of sparse LiDAR point cloud for depth acquisition. However, as pixel resolution was much greater than LiDAR point cloud one, many pixels were not assigned the depth value, thus extra interpolation was needed to

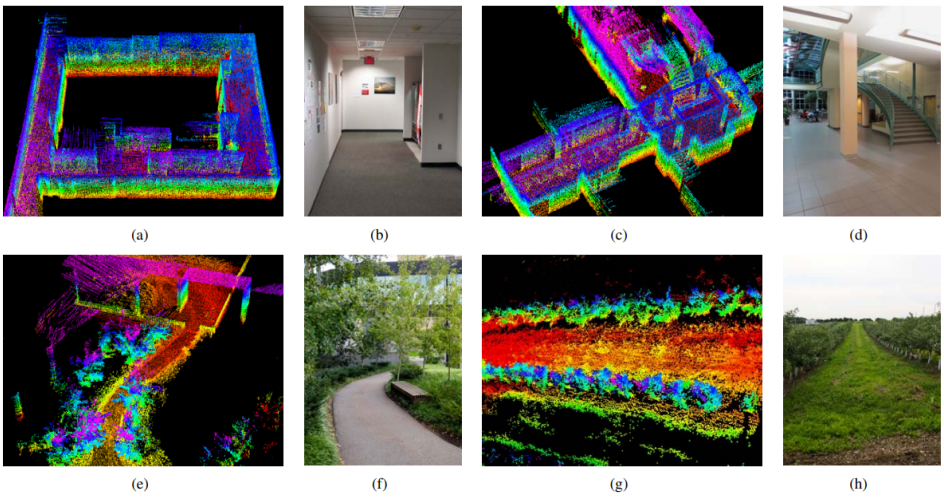


Figure 1.8 – The LOAM algorithm is tested on both indoor and outdoor scenarios for dense scene reconstruction based on the estimated poses [15]

make up the missing values. In many cases, LiDAR scan-matching is used for local motion estimation and visual hint is utilized for loop closure validation. The accuracy of LiDAR based localization was improved in [19], with visual feature aided loop detection to reduce the accumulated drift. In [20], the visual keyframes were utilized to assist the laser-based slam to perform local and global bundle adjustments. Furthermore, the LiDAR scan-to-scan matching can be improved using the initial guess from visual estimation as demonstrated in [21]. There are also many works which coupled both LiDAR and visual state estimation process together. Zhang et al. [22] designed V-LOAM pipeline which used high frequency vision based odometry as the motion prior and corrected with high precision, low frequency lidar scan matching estimation afterwards. The framework in [23] did not rely on visual estimation as the motion initial guess for lidar odometry. They took in both visual and LiDAR measurements, stacking and minimizing both modalities' residuals during the optimization phase. However, as mentioned beforehand, they did not consider the uncertainty during state estimation process, which may cause the overconfident estimation prone to certain sensor modality.

1.3 Thesis Contribution and Outline

The main contributions of this thesis are three-fold.

1. **Enhanced Vision-based Perception and Localization.** The mono-vision system can efficiently localize the ego-vehicle, while creating a map of the surroundings in static scenarios. And the scale metric can be recovered trivially with the stereo-vision system. However, the vision-based estimation tends to have poor performance with the moving objects in the scene, which is quite common in real traffic scenarios. In order to segment the moving objects from the background, pure geometric modeling is insufficient and semantic information needs to be associated to the geometric entities. A semantic-guided method is thus proposed to address this issue, which ensures the data association process is not biased due to the impact of moving objects presence. Moreover, the backward covariance propagation is leveraged to transform the uncertainty from the measurement space to the estimation space. And the estimation uncertainty paves the way to adaptive sensor fusion.

2. **Enhanced LiDAR-based Perception and Localization.** The LiDAR is capable of accurate range sensing, which provides the basis for scale-aware pose estimation and scene reconstruction. In order to overcome the weakness of individual direct or feature-based LiDAR scan matching, a semi-direct LiDAR point cloud registration method is proposed. The adaptive initialization strategy takes advantage of the constant-velocity motion model, and it reduces the possibility that the initial guess position is in the vicinity of local minima. Besides, the uncertainty of LiDAR-based ego-motion estimation is derived with inconsistency indicator. The pose uncertainty estimation helps to bound the pose error within a known confidence interval. On this basis, the main factors which degrade LiDAR scan alignment performances are analyzed and eased by the pose graph-based optimization. It is demonstrated that ground points and dynamic objects such as vehicles or pedestrians are the main causes of pose estimation accuracy decrease. It is also noticed that significant errors frequently occur near the road intersections and in highway scenarios, where it is more likely to come across the dynamic vehicles and the geometric information is not adequate for reliable ego-motion estimation. Moreover, the LiDAR-based object detection problem is investigated with both unsupervised clustering-based and supervised learning-based methods, which further empowers the onboard perception system. And it is shown that the CenterPoint neural network outperforms the state-of-the-art for the objects orientation prediction in real traffic scenarios.

3. **Adaptive LiDAR-Vision Fusion-based Perception and Localization.** For the autonomous vehicle, it is not reliable to perceive the environment with single sensing modality. In this thesis, the complementary visual and range sensors are adaptively combined for robust localization and consistent scene mapping. To be more specific, the LiDAR sensor is employed for 3D real-dimension object detection and the visual features are leveraged for the objects' state of motion (static/dynamic) estimation. With the dynamic portions of the scene identified, the data association problem is easy to solve and the scene mapping will be more consistent. Additionally, a loosely coupled vision-LiDAR odometry is proposed considering the uncertainty of each sensor estimation. Covariance intersection filtering ensures that the uncertainty of pose estimation does not expand after the sensor fusion, which efficiently filters out the unstable estimation. The fused pose will better register the lidar map points and the multi-level voxel scan-to-map matching is adopted to reduce the frame-to-frame estimation drift.

Some of the thesis results have been published in international conference proceedings or have been submitted to the international journal for peer-review.

Peer-Reviewed International Conferences

— Chen, S., Sun, H. and Frémont, V., A Semantic-Guided LiDAR-Vision Fusion Approach for Moving Objects Segmentation and State Estimation. 2022 IEEE International Intelligent Transportation Systems Conference (ITSC 2022), Oct 2022, Macau, China.

— Chen, S., Sun, H. and Frémont, V., Mono-Vision based Moving Object Detection using Semantic-Guided RANSAC, 2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2021), Sep 2021, Karlsruhe, Germany.

— Chen, S. and Frémont, V., A Loosely Coupled Vision-LiDAR Odometry using Covariance Intersection Filtering. 2021 IEEE Intelligent Vehicles Symposium (IV 2021), Jul 2021, Nagoya, Japan.

Peer-Reviewed International Journals

— Chen, S. and Frémont, V., Uncertainty-Aware Semi-Direct LiDAR Scan Matching for Ground Vehicles Positioning and Scene Mapping. Submitted to IEEE Sensors

Journal (under review).

The following chapters of this thesis is organized as follows. To begin with, in Chapter 2, vision-based perception and ego-motion estimation are explored. Moving objects are efficiently segmented from the background with the semantic-guided RANSAC algorithm. Moreover, the pose estimation uncertainty modeling with the backward covariance propagation is demonstrated. Then, the investigation of LiDAR-based perception and localization is conducted in Chapter 3. A semi-direct LiDAR scan matching approach is proposed for robust ego-motion estimation. And the LiDAR-based pose estimation uncertainty is deduced from the inconsistency indicator, from which the main factors to degrade the laex ser scan alignment can be analyzed. Moreover, to facility LiDAR-based scene understanding, both clustering-based and learning-based 3D objects detection are implemented and discussed. Finally, in Chapter 4, the LiDAR-Vision fusion-based approach is developed for better scene perception and localization. The proposed algorithms are tested with real world datasets and promising benchmarking results are obtained. The achieved leading results reflect the advantages of LiDAR-Vision measurements fusion and geometric-semantic information integration in complex traffic scenarios.

Chapter 2

Vision-based Perception and State Estimation

It is fundamental for the autonomous vehicles to obtain an accurate scene perception and precise self-localization, which will facilitate their secure operation in various unknown environments. Hereby, the vision-based perception stands for objects detection on the image plane and state estimation includes both the ego-vehicle poses and map points location. Vision-based Advanced Driver Assistance Systems (ADAS) are widely employed onboard nowadays to acquire the information of the surroundings and perform high-level tasks such as autonomous parking, adaptive cruise control, obstacle avoidance, etc. The visual sensors like cameras are just like human eyes, which receive the passive light and transform the 3D world into the 2D images via projection. Due to their affordable cost and compact size, the cameras are usually prioritized to be integrated into the intelligent vehicle perception system. The cameras are versatile and can capture the rich color and texture information of the surroundings with low latency. Based on the visual measurements, various distinctive features can be extracted and matched for ego-motion estimation. Moreover, with the learning-based methods, meaningful objects and semantic labels can be predicted in urban traffic scenarios with the geometric features encoding and decoding. Thus, the visual sensors play a key role for the autonomous driving system to identify the scene components such as pedestrians, cyclists, vehicles, lane markings, traffic signs, etc.

The perception system that incorporates only a single camera is called mono-vision perception system. The mono-vision system can efficiently capture the 3D environment and estimate the camera pose with an up-to-scale factor due to the depth information loss. Meanwhile, the existence of moving objects in the scene is quite common, and it will degrade the ego-motion estimation. Segmenting these moving objects is essential for robust Structure-from-Motion (SfM) process. The semantic-guided approach will be detailed in this section to leverage the contextual and geometric information together

for moving objects segmentation. In order to solve the scale ambiguity of the mono-vision system, the stereo-vision system is proposed, where the pixel depth value can be inferred from the left-right images matching and disparity calculation. Then the 3D scene reconstruction task can be implemented with the depth information, which is the prerequisite for the further map-based localization or other interaction. Moreover, due to the measurement noise, the estimation results are not accurate or perfect. The ego-motion uncertainty modeling will also be discussed, which propagates the uncertainty from the measurement space to the estimation space based on the backward propagation method.

2.1 Mono-Vision based Structure-from-Motion

Structure from Motion (SfM) is the process to incrementally compute the camera poses and to recover the 3D scene structure with the input of sequential 2D images. The 2D images captured by the pinhole monocular camera are essentially the perspective projection of the 3D world (see Fig. 2.1). The image projection consists of the rays

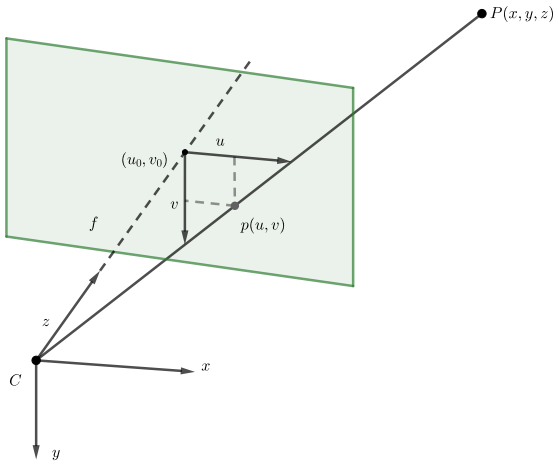


Figure 2.1 – Scheme of the pinhole camera perspective projection model

reflected from the scene components, which lie on the image plane with the distance of focal length from the camera optical center. Let $\mathbf{P} = [x, y, z]^T$ be a 3D point expressed in the camera frame $\mathcal{F}_C$ and $\mathbf{p} = [u, v]^T$ be the pixel coordinates on the image plane. The

3D-2D pinhole camera perspective projection can be described as follows:

$$z\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{KP} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{2.1}$$

where z encodes the point depth value, and f_x, f_y, u_0, v_0 are the camera intrinsic parameters. The camera intrinsic parameters can be estimated with Zhang's method [24] when the planar checkerboard dimension and structure are known, see Fig. 2.2 and Fig. 2.3. In order to have a non-biased calibration result, the checkerboard pattern needs to be placed on a planar surface and be captured from different position and orientation. According

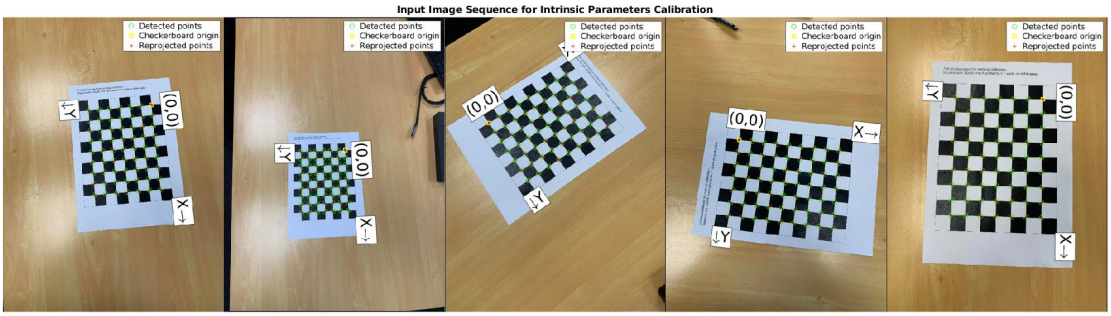


Figure 2.2 – The input image sequence for the camera intrinsic and extrinsic parameters calibration

to whether the visual features are detected and tracked for the camera pose estimation, the monocular SfM problem can be solved with the appearance-based or feature-based methods. The appearance-based methods [25] [26] operate directly on image pixels and leverage the pixel intensity information for the ego-motion estimation. Instead of direct photometric error minimization, the feature-based approaches [9] [12] track the distinctive and repeatable features across frames and aim to minimize the reprojection errors. The corner features proposed in [27] [28] [29] can be extracted trivially on the image plane and they are more invariant to the illumination changes than the simple gray-scale pixel values. To account for the rotation, scale and lighting condition variations during the viewpoint changes, more and more handcrafted visual features such as Scale-Invariant Feature Transform (SIFT) [30], Speeded Up Robust Features (SURF) [31], Oriented fast and Rotated Brief (ORB) [10] are proposed. Among them, the ORB feature is widely

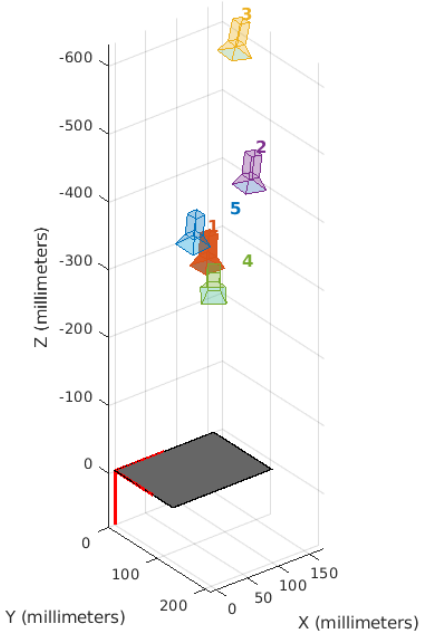


Figure 2.3 – The estimated camera poses during the camera calibration process

adopted due to its real-time performance of feature detection and matching, see Fig. 2.4. Meanwhile, the rotated Binary Robust Independent Elementary Features (BRIEF) [32] descriptor is embedded in the ORB feature, which allows for the large inter-frame movement feature matching according to the descriptor similarity.

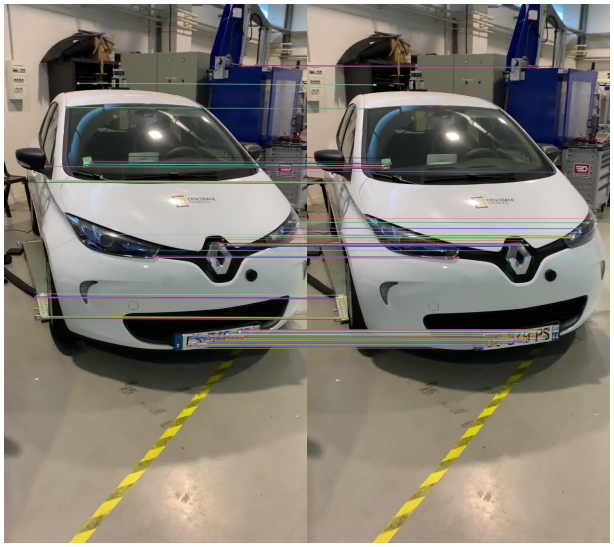


Figure 2.4 – The ORB features are detected and matched on two adjacent ZOE images

Given the matched 2D feature points on the image plane, the up-to-scale camera motion can be estimated with the epipolar constraints [33], see Fig. 2.5. Suppose that $\mathbf{p}_1$

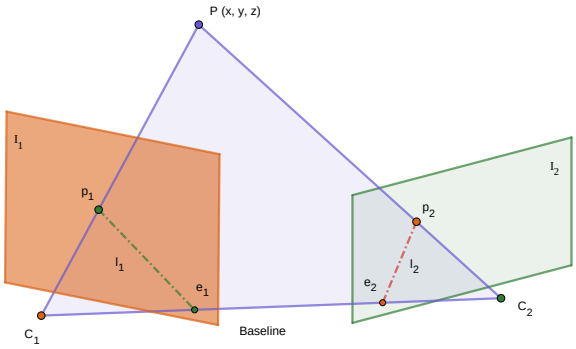


Figure 2.5 – Scheme of the epipolar geometry

and $\mathbf{p}_2$ are the projection of the world point $\mathbf{P}$ on the subsequent image planes, then the

epipolar constraint can be written as:

$$\mathbf{p}_2^T \mathbf{K}^{-\mathrm{T}} \mathbf{t}^{\wedge} \mathbf{R} \mathbf{K}^{-1} \mathbf{p}_1 = 0 \tag{2.2}$$

where the $\mathbf{K}$, $\mathbf{t}$ and $\mathbf{R}$ represent the camera intrinsic matrix, camera translation and camera rotation respectively. It can be seen that both camera translation and rotation are encoded in the epipolar equation Eq. 2.2, the essential matrix $\mathbf{E}$ and fundamental matrix $\mathbf{F}$ can be defined as follows:

$$\mathbf{E} = \mathbf{t}^{\wedge} \mathbf{R}, \quad \mathbf{F} = \mathbf{K}^{-\mathrm{T}} \mathbf{E} \mathbf{K}^{-1}, \quad \mathbf{x}_2^T \mathbf{E} \mathbf{x}_1 = \mathbf{p}_2^T \mathbf{F} \mathbf{p}_1 = 0 \tag{2.3}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the normalized coordinates. In reality, the matched feature points will be more than 8 pairs. In this situation, the Random Sample Consensus (RANSAC) [6] algorithm needs to be applied to filter out the spurious measurements and estimate the model with the most inliers. The principle of RANSAC is to compute the model hypotheses from randomly selected samples and verify the hypotheses with the rest of the measurement set. To be more specific, for each iteration, eight points are randomly sampled from the entire matched points set. Then the hypothesis fundamental matrix can be estimated with the eight-point algorithm [34]. For the rest of the paired points, their epipolar distances are computed and summed up for hypothesis verification. At the end of the iterations, the fundamental matrix with the least summed epipolar distances is chosen as the estimation result. Meanwhile, the matched points can be classified as inliers or outliers according to the predefined threshold. Due to the depth information loss, the epipolar geometry can only estimate an up-to-scale camera translation $\mathbf{t}$. By convention, $\mathbf{t}$ is usually normalized with the norm of 1.

After estimating the camera motion, the observed feature points on the image plane can be triangulated to obtain the 3D spatial position of the points, see Fig. 2.6. Triangulation means that the pixel depth value can be determined if the points (landmarks) are observed from two image frames with the relative pose of $\mathbf{R}$ and $\mathbf{t}$. Ideally, the two rays $\overrightarrow{\mathbf{C_1 p_1}}$ and $\overrightarrow{\mathbf{C_2 p_2}}$ will intersect at the 3D map point $\mathbf{P}$. However, due to the measurement noise, the two rays will not intersect perfectly. In this case, the least-square method can be deployed to minimize the intersection discrepancy. According to the epipolar constraints, the normalized coordinates $\mathbf{x}_1$ and $\mathbf{x}_2$ will follow:

$$s_2 \mathbf{x}_2 = s_1 \mathbf{R} \mathbf{x}_1 + \mathbf{t} \tag{2.4}$$

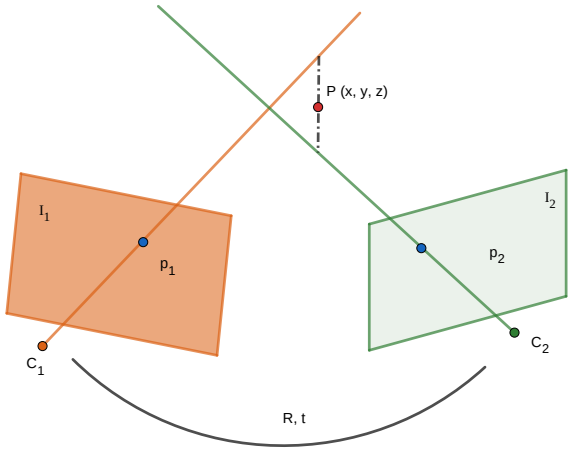


Figure 2.6 – Scheme of the feature point triangulation

where s_1 and s_2 are depth of the observed feature points. If $\mathbf{x}_2^\wedge$ is multiplied on the both sides of the equation:

$$s_2\mathbf{x}_2^\wedge\mathbf{x}_2 = 0 = s_1\mathbf{x}_2^\wedge\mathbf{R}\mathbf{x}_1 + \mathbf{x}_2^\wedge\mathbf{t} \tag{2.5}$$

where $\mathbf{x}_2^\wedge$ is the cross product skew symmetric matrix. Then, the depth value can be trivially solved. The deviation distances from the triangulated 3D points to the camera optical rays quantify the triangulation quality, and the 3D points which exhibit large uncertainty will be dropped out. After the initialization phase of the monocular SfM, incrementally, the camera poses and corresponding 3D map points can be recovered. Since the absolute scale can not be obtained from the mono-vision system, for consecutive frames, the relative scale ratio r can be calculated as:

$$r = \frac{\left\| {}^{\mathcal{F}_{C_{t-1}}}\mathbf{P}_i - {}^{\mathcal{F}_{C_{t-1}}}\mathbf{P}_j \right\|}{\left\| {}^{\mathcal{F}_{C_t}}\mathbf{P}_i - {}^{\mathcal{F}_{C_t}}\mathbf{P}_j \right\|} \tag{2.6}$$

where $\mathbf{P}_i$ and $\mathbf{P}_j$ are the 3D co-visible map points in the frame $\mathcal{F}_{C_{t-1}}$ and frame $\mathcal{F}_{C_t}$. In order to reduce the motion estimation drift, windowed Bundle Adjustment (BA) [35] optimization can be opted to adjust the 3D map points and camera poses jointly, see Fig. 2.7. The BA reprojection residual can be expressed as follows:

$$\arg\min_{\mathbf{P}^i,\mathbf{C}_k} \sum_{i,k} \left\| \mathbf{p}_k^i - f\left(\mathbf{P}^i, \mathbf{C}_k\right) \right\|^2 \tag{2.7}$$

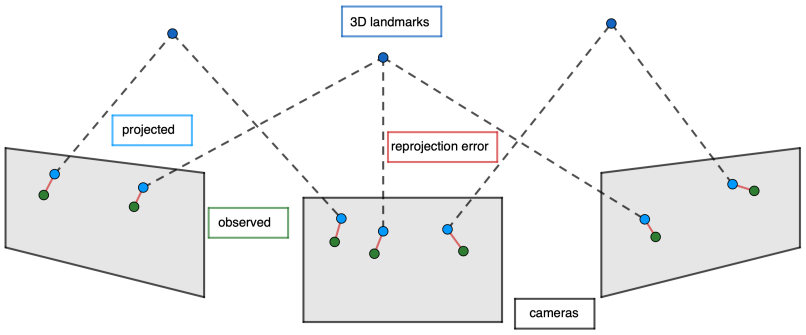


Figure 2.7 – Scheme of the bundle adjustment reprojection error minimization

where $\mathbf{p}_k^i$ is the image point of the 3D landmark $\mathbf{P}^i$ measured in the k^{th} image and $f(\mathbf{P}^i, \mathbf{C}_k)$ is reprojected image point with the estimated camera pose $\mathbf{C}_k$ and estimated map point $\mathbf{P}^i$. The windowed bundle adjustment is essentially a nonlinear optimization problem, and can be efficiently solved with the Levenberg-Marquardt method, owing to the sparsity feature of the BA problem. The sparsity feature of BA means that the estimated camera poses and 3D landmarks are only related with the co-visibility measurements, which makes the Jacobian matrix sparse and easy to solve. During the nonlinear optimization, Lie algebra [36] is leveraged to describe the camera poses $\mathbf{C}_k$, which makes the Jacobian derivatives much simpler without the need to consider the rotation matrix orthogonality constraints. The local bundle adjustment refines the camera poses (rendered in red) and provides the sparse feature-based scene reconstruction, as seen in Fig .2.8. The full sequence image samples can be accessed from the link. For solving the depth scale ambiguity, the stereo vision system can be employed or the range sensor such as LiDAR can be integrated into the perception system, which will be discussed in the following Section 2.3 and Section 4.2.

2.2 Mono-Vision based Moving Object Segmentation

Vision-based traffic scene understanding is a complex yet indispensable task for the perception of autonomous vehicles. Typically, vision-based Moving Object Segmentation (MOS) is fundamental for high-level tasks such as obstacle avoidance in dynamic and uncertain environments. Identifying the dynamic objects also plays an important role in the vision based ego-motion estimation problem which usually has the assumption of

(a) The captured ZOE image sequence samples

(b) The Structure from Motion result with the ZOE image sequence

Figure 2.8 – The scheme for Stucture from Motion pipeline to reconstruct the scene

static surroundings. Being able to recognize movable objects (cars, pedestrians, cyclists) and to obtain their states (stationary, non-stationary) can facilitate the safety of the autonomous vehicle operation.

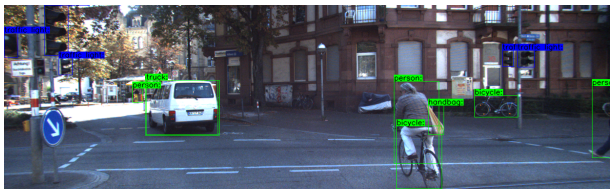
Before the wide application of deep learning-based methods for scene perception, local sliding-window feature-based approaches based on Scale-Invariant Feature Transforms

(SIFT) [37] and Histogram of Oriented Gradients (HOG) [38] are commonly adopted for objects detection on the image plane. Lowe [37] demonstrated that the extracted local visual features can be leveraged to identify the potential objects with the near-neighbor indexing. In [38], the HOG feature vectors are calculated and fed in the Support Vector Machine (SVM) for object/non-object classification. And the conventional non-maximum suppression technique is then used to filter out the detection false alarms. However, it remains challenging for the feature-based methods to identify the unseen objects in the scene, which limits their application. In the machine learning era, these difficulties are well addressed with the deep learning-based approaches that greatly outperform the traditional feature-based methods. Generally, the learning-based object detection methods can be categorized as two types: the two-stage region-based methods [39] [40] [41] and one-stage region proposal-free methods [42] [43]. Regions with Convolutional Neural Networks (R-CNN) in [39] adopts a selective search algorithm to generate up to 2000 region proposals. Then the CNN architecture is leveraged in these Regions of Interest (ROI) to extract dense features, and the feature vector will be fed into the SVM classifier to predict the presence of objects. Additionally, the post-processing is implemented to refine the bounding box dimensions and eliminate duplicate detection results. Instead of applying the CNN separately to different ROIs, Fast R-CNN [40] directly generates a convolutional feature map through the single image forward pass. From the convolutional feature map, the bounding box location and class can be regressed together after the ROI pooling layer to improve the efficiency. Furthermore, faster R-CNN is proposed in [41] which unifies the ROI proposal, feature extraction and object classification within the CNN architecture to accelerate the object detection process. A Region Proposal Network (RPN) is used to predict the ROIs, which is much more efficient than the previous selective search algorithm. Unlike sliding window and region proposal-based methods, the single shot detector You Only Look Once (YOLO) [42] encodes global contextual information of the entire image at the first sight and converts object detection into a single regression problem. Owing to this unified design, the real-time (155 fps) bounding box prediction and classification can be achieved. The YOLO network [42] is further enhanced in YOLO9000 [43], where the detection performance is optimized and up to 9000 categories objects can be identified. The YOLO architecture cuts the image into different segments and outputs instance-level bounding boxes which are weighted by the prediction scores. Though the YOLO network is able to predict the existence of objects and their semantic class in the scene, the objects state of motion prediction is still challenging for the end-to-end neural-based approaches.

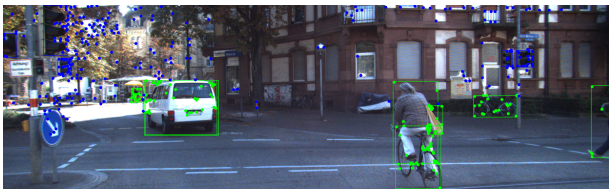
Specifically, the object state of motion remains unknown if its semantic label is not definitely static. In Microsoft COCO dataset [44], definitely static (non-movable) objects are listed as traffic light, fire hydrant, stop sign, parking meter, bench and potted plant. The rest are classified as movable objects such as pedestrians, bicycles, or vehicles, which need further information to solve the ambiguities for their states of motion.

Substantial research work has been devoted to the domain of moving objects segmentation in recent years. Background subtraction [45] approach is widely applied to handle the MOS problem when image sequences are acquired from a static camera. However, for a moving camera, this approach cannot be directly utilized without additional constraints imposed. Because of the vehicle ego-motion, the object-motion and ego-motion are coupled together which makes the background subtraction non-trivial. In order to decouple and compensate for the ego-motion, the epipolar geometry [45] is commonly adopted for ego-motion estimation across two consecutive frames. Unfortunately, sparse feature-based state estimation may be unstable when the non-static feature points are chosen and incorporated in the estimation process. By convention, dynamic objects are regarded as outliers and a random sample consensus (RANSAC) [6] method is often applied to filter them out. However, this strategy fails to operate when the dynamic objects turn out to be the dominant components in the scene. Thus, effective moving objects segmentation in a complex scene remains a critical issue to be solved for the perception of autonomous vehicles. In [46], the challenge of estimating a vehicle's ego motion as well as the movements of dynamic objects at the same time is addressed based on projective factorization of the multiple-trajectory matrix. Stereo-vision based moving object segmentation methods are proposed in [47] [48] [49], where the motion likelihood of every pixel is calculated given the approximated ego-motion uncertainty and U-disparity map is built to characterize on-road obstacles. Color and depth hints are leveraged in the graph-cut framework for connected regions (moving objects) extraction. In [50], a Bayesian framework is applied to generate a probability value for each pixel, either being static or dynamic, according to the epipolar and focus of expansion constraints. The framework enables the system to segment moving objects with degenerate motion due to the Flow Vector Bound (FVB) constraint attached. With the advancement of learning-based methods, an unsupervised adversarial contextual model is proposed in [51] to segment dynamic objects in the image frame. The contextual information of the surroundings is fed for the neural network training to infer the optical flow in specific regions, meanwhile another network formats the context as uninformative as possible since the optical flow of a moving object is uncor-

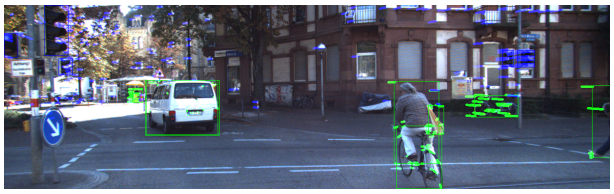
related with the background. The term of moving objectness is introduced in [52], which represents the possibility that they belong to moving objects. Several prediction are firstly proposed using multiple figure-ground segmentations and then the proposals are ranked with the moving objectness criteria to identify moving objects. In [53], Neural-Guided RANSAC is applied to a wide range of computer vision tasks such as fundamental matrix estimation, horizon line estimation and camera re-localization. Inspired by this, a novel semantic-guided RANSAC algorithm is formulated in this thesis to reject instance-level outliers which helps to discriminate truly moving objects from stationary ones. Different from differentiable RANSAC in [53], the proposed two-stage approach (semantic prediction and geometric validation) is more flexible to add constraints to detect the objects with degenerate motion regardless of the ego-motion variation, without modifying or retraining the existing neural networks.



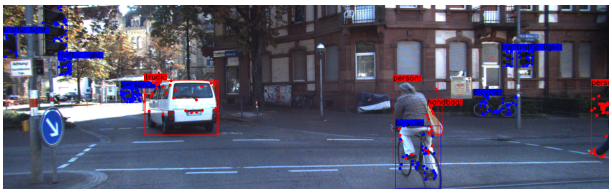
(a) Get the bounding boxes with semantic labels using yolo v4 detector



(b) Extract the Shi-Tomasi corner feature points, rendering in green for movable objects, rendering in blue for non-movable objects and background points



(c) Estimate the Fundamental matrix with the feature correspondences rendering in blue (pyramidal Kanade-Lucas-Tomasi tracker)



(d) Render the moving bounding box in red when the proportion of pixel-level outliers in the bounding box is over the threshold of 0.5

Figure 2.9 – Overview of the proposed semantic-guided RANSAC for moving objects segmentation. Blue, red, green colors stand for stationary, non-stationary and unknown objects respectively in KITTI dataset

In Fig. 2.9, it is shown that the proposed framework starts with a YOLO object detection module. Objects with static semantic labels such as traffic lights are directly classified as stationary. However, movable objects with the labels such as person, bicycle and car need further information to make the inference. Thus, Shi-Tomasi corner points [28] are extracted from the image and iteratively tracked using Lucas-Kanade optical flow.

Feature points which belong to the static objects and background are utilized to estimate the fundamental matrix. Semantic-guided RANSAC takes full advantage of instance-level semantic segmentation and enables the fusion of semantic labels and geometric constraints for moving object detection. Combining semantic and geometric cues results in accurate moving object segmentation by checking the residual value of the epipolar constraint and flow vector bound for all suspicious points lying in the movable bounding boxes. Instead of training an end-to-end fashion neural network which outputs object existence and its state, a two-stage approach is taken in this thesis. The YOLO network output provides good semantic prior to predict the existence of the objects and then semantic-guided ransac decides the state of the objects based on the epipolar geometry and flow vector bound constraint.

The well-known Kanade–Lucas–Tomasi (KLT) [54] tracker leverages spatial intensity cues to guide the search for the corresponding features across two frames. In order to deal with large camera motion across frames, a pyramidal KLT tracker is implemented to allow for tracking points with large displacements between frames. Moreover, semantic label consistency and forward-backward flow consistency constraints are added into feature points tracking process to reduce the occurrence of mismatches due to occluded pixels and pixels with strong illumination changes. Pyramidal KLT tracker can be applied to get the pair of matched feature points $(\mathbf{p}_1, \mathbf{p}_2)$ across frames, $\mathbf{p}_2 = KLT_{forward}(\mathbf{p}_1)$. Then the optical flow propagates backward to get the estimated initial feature point $\hat{\mathbf{p}}_1$, $\hat{\mathbf{p}}_1 = KLT_{backward}(\mathbf{p}_2)$. The forward-backward constraint is imposed to compute the euclidean distance $dist(\hat{\mathbf{p}}_1, \mathbf{p}_1)$ for matched points, and this metric is used to discard potentially erroneous feature matches when their discrepancy is over 2 pixels.

$$Status(\mathbf{p_1}, \mathbf{p_2}) = \begin{cases} valid & dist(\hat{\mathbf{p}}_1, \mathbf{p}_1) < 2 \\ invalid & dist(\hat{\mathbf{p}}_1, \mathbf{p}_1) > 2 \end{cases} \tag{2.8}$$

$$Status(\boldsymbol{p_1}, \boldsymbol{p_2}) = \begin{cases} valid & label(\boldsymbol{p}_1) = label(\boldsymbol{p}_2) \\ invalid & label(\boldsymbol{p}_1) \neq label(\boldsymbol{p}_2) \end{cases} \tag{2.9}$$

RANSAC [6] is an iterative method to estimate the underlying model parameters which meanwhile divides the input data into inliers and outliers. The main limitation of RANSAC is that, when large number of outliers are incorporated in the dataset, biased estimation output may be provided due to the limits of iteration times. Due to the increase of outliers ratio, RANSAC needs exponentially more iterations to reach the point with a

outlier-free subset found, see Fig. 2.10. The expected number of iterations r to reach a certain probability p with a minimal outlier-free subset found is

$$r = \frac{\log(1-p)}{\log\left(1-w^N\right)} \tag{2.10}$$

where w is the fraction of inliers and N is the minimum number of samples needed for model estimation which should be eight pairs [45] of matching points for fundamental matrix estimation in our case. Higher inlier fraction is preferred since it helps to incorporate more correct correspondences in the consensus set and fewer iterations are needed get obtain the model parameters.

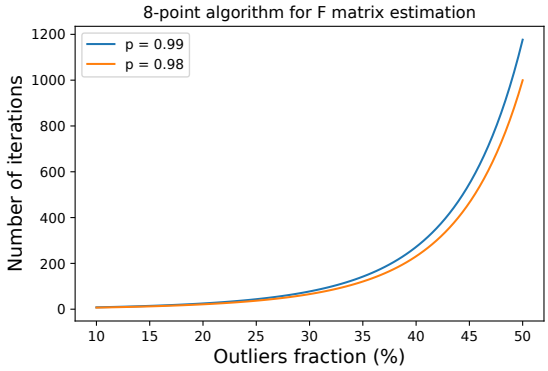


Figure 2.10 – RANSAC number of iterations for 8-point fundamental matrix estimation

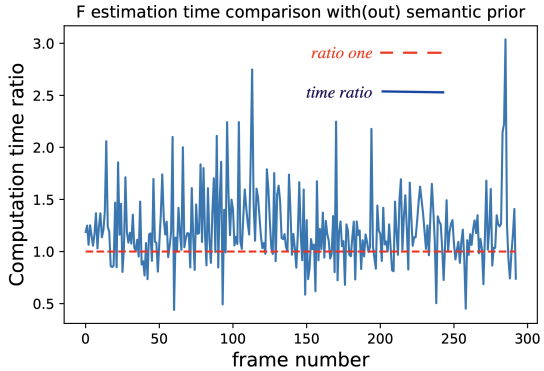


Figure 2.11 – The ratio of fundamental matrix estimation time with/out semantic prior. The semantic prior can reject the outliers and accelerate the fundamental matrix estimation convergence

The proposed semantic-guided fundamental matrix estimation makes use of semantic

priors to guide the model fitting, which facilitates obtaining the outlier-free minimal subset. Static feature points belonging to background and objects with non-movable semantic labels such as traffic lights and traffic signs have higher priority than movable objects such as pedestrians and vehicles to be utilized for fundamental matrix estimation. Moreover, the semantic bounding box from YOLO v4 output whose objectness score is lower than a threshold of 0.2 is suppressed and not taken into account. The semantic prior ultimately increases the fraction of inliers in the tentative pairs set and as a result, fundamental matrix estimation is more well conditioned which requires much less number iterations to converge (see Fig. 2.11), where the frames are taken from KITTI raw data 2011_09_26_*drive*_0056 sequence.

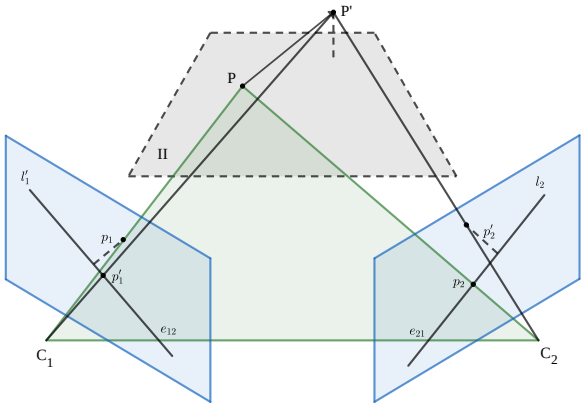
Considering pairs of matched points belonging to the background and static objects, the fundamental matrix is robustly estimated with the 8-point algorithm [34]. And given a pair of matched points $(\mathbf{p}_n^i, \mathbf{p}_{n+1}^i)$ lying in the movable object, geometric constraint can be leveraged to distinguish the truly dynamic objects from the static ones. Fundamental matrix maps the point $\mathbf{p}_n^i$ to its corresponding epipolar line $\mathbf{l}_{n+1}^i$ as $\mathbf{l}_{n+1}^i \sim \mathbf{F}\mathbf{p}_n^i$ across two frames, where $\sim$ represents an up-to-scale equality. Then it is possible to calculate the epipolar geometry residual $r_{\boldsymbol{F}}$ for matched points to implement outlier rejection based on point-to-line distance d_{p2l} in the image.

$$r_{\boldsymbol{F}} = \max\{d_{p2l}\left(\mathbf{p}_n^i, \mathbf{l}_n^i\right), d_{p2l}\left(\mathbf{p}_{n+1}^i, \mathbf{l}_{n+1}^i\right)\} \tag{2.11}$$

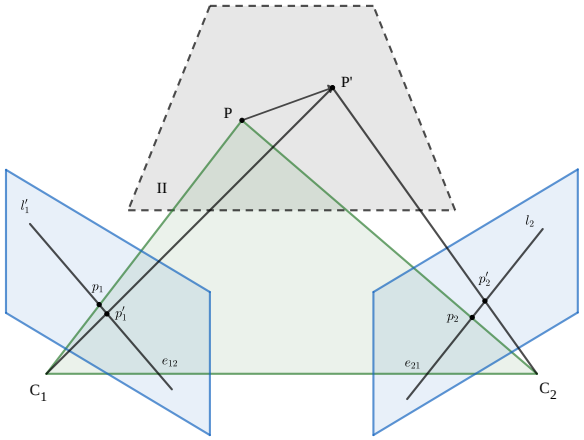
where the point-to-line distance d_{p2l} from the point $\mathbf{p}_n^i = (u_n^i, v_n^i)$ to line $\mathbf{l}_n^i$ with reduced coefficients $[a_n^i, b_n^i, c_n^i]^T$ is calculated as:

$$d_{p2l} = \frac{a_n^i u_n^i + b_n^i v_n^i + c_n^i}{\sqrt{(a_n^i)^2 + (b_n^i)^2}} \tag{2.12}$$

However, when a 3D point in the scene appears on the epipolar plane which is constructed by the point itself the camera center in the previous and current frames, the perspective projection of the moving point always stays on the corresponding epipolar line. In this case, null epipolar residual does not represent that the point is static, see Fig. 2.12. So the epipolar constraint is not capable to detect such moving points with degenerate motion. Thus, another constraint using Flow Vector Bound (FVB) [50] is additionally imposed to find the bound of parallax range $[d_{min}^{FVB}, d_{max}^{FVB}]$ for static and background points in the scene with Eq. 2.13. Given images captured from a pinhole camera, pixel-wise displacement d^i



(a) The epipolar geometry constraint works well in non-degenerate case



(b) The epipolar geometry constraint fails in degenerate case

Figure 2.12 – The scheme for the degenrate motion illustration

for the feature point $\mathbf{p}_n^i = (u_n^i, v_n^i)$ which has the depth value z can be obtained with the equation:

$$\begin{aligned} \mathbf{p}_{n+1}^i - \mathbf{K}\mathbf{R}\mathbf{K}^{-1}\mathbf{p}_n^i &= \tfrac{1}{z}\mathbf{K}\mathbf{t} \\ d^i &= \left|\mathbf{p}_{n+1}^i - \mathbf{K}\mathbf{R}\mathbf{K}^{-1}\mathbf{p}_n^i\right| \end{aligned} \tag{2.13}$$

where $\mathbf{K}$, $\mathbf{R}$ and $\mathbf{t}$ are the camera intrinsics, rotation and translation from timestamp n to $n+1$ respectively. Any point with a parallax value d^i which is not in the range $[d_{min}^{FVB}, d_{max}^{FVB}]$ will be also set as an outlier and rejected. Then, the ambiguous movable object can be classified as a truly dynamic (static) object if there are more than 8 feature points lying on the object and the proportion of pixel-level outliers is above (below) the threshold of 0.5. If the number of feature points inside the bounding box is less than

8 (minimum number for independent fundamental matrix estimation), the state of the object is set as unknown and waiting for further information to make the decision. In Section 4.1, a more sophisticated σ-rule based probabilistic method will be developed to reduce the number of tuning hyper-parameters. Moreover it needs to be noted that epipolar geometry constraint only works for the moving camera mounted on the vehicle. When the camera does not move, the epipolar line can not be defined. In this situation, FVB constraint complements the epipolar geometry constraint to detect moving objects in the scene.

Algorithm 1 Semantic-Guided Instance-level Outlier Rejection

Input: Corresponding feature points in two consecutive frames

Output: Segmented moving bounding boxes in the scene frame

1: ▶ Extract the background and static feature points in the scene by excluding the feature points with movable semantic labels
2: ▶ Apply the 8-point algorithm to estimate $\boldsymbol{F}$ with the static and background feature points
3: ▶ Check how well $\boldsymbol{F}$ matches feature points in the bounding boxes with movable labels using Eq. 2.11
4: ▶ Impose FVB constraint to detect feature points on objects with degenerate motion using Eq. 2.13
5: ▶ Determine the movable object as the truly dynamic object if there are enough feature points on the object and the proportion of pixel-level feature point outliers is above the threshold of 0.5

The KITTI dataset [55] contains image sequences recorded in urban and highway environments. In the category of raw data, 2D bounding boxes tracklets of moving objects are provided for several sequences. Our system is evaluated at the bounding box level with the metrics of Precision and F-score defined as:

$$P = \frac{t_p}{t_p + f_p} \ , \ F = \frac{2t_p}{2t_p + f_p + f_n} \tag{2.14}$$

with t_p, f_p and f_n represent true positive detection, false positive misdetection and false negative alarm successively. In order to highlight the advantage of our proposed approach which fuses semantic and geometric information, the method presented in [48] and [49] which uses stereo-vision without considering semantic clues are chosen as the base-

Figure 2.13 – Flow vector bound constraint for detecting moving objects with degenerate motion with the frame from KITTI raw data sequence 05

(a) YOLO network wrongly classifies reflections in the mirror as cars with high confidence

(b) States of cars in the mirror are set as unknown (green) in our pipeline due to lack of consistently matched feature points for decision-making

Figure 2.14 – Result analysis for false alarms due to mirror reflection with the frame from KITTI raw data sequence 71

line. Four typical heavy traffic scene sequences are selected to evaluate our system. Tab. 2.1 illustrates the quantitative results for the comparison purpose. From Tab. 2.1, it is shown that the precision of moving object detection has been greatly improved due to the semantic cues involvement. Taking the semantic information into account increases the true positive detection rate and decreases the false negative alarm rate. Moreover, with RTX 2070 GPU acceleration, the neural-based semantic inference can run at the speed of 30FPS. Then, with the efficient CPU-based OpenCV library [56] for geometric validation and instance-level outlier rejection, the cascaded implementation can achieve real-time performance at 10FPS. It is superior to the approach in [48] and which takes more than 0.2 second to estimate the ego-motion along with its uncertainty for each single frame. In the proposed method, the ego-motion is implicitly integrated in the epipolar geometry point-to-line residuals and the sparse feature points optical flow ensures the efficiency of the whole pipeline. Fig. 2.13 demonstrates that, the FVB constraint effectively helps to segment moving objects with degenerate motion. At the same time, the FVB constraint also applies when the ego-motion is null, see the video presentation, which makes our system robust against the ego-motion variation. And it is presented in Fig. 2.14 that, the false alarms due to mirror reflection are set as unknown states because of the minimum

feature number constraint imposed. Compared with the end-to-end moving object detection approach, our two-stage method is more flexible to add constraints without modifying or retraining the existing neural networks. Despite these advantages, our framework also has some drawbacks. It does not perform very well in some certain scenario where the objects are far from the ego-vehicle. In this situation, they appear to be very small and there are not enough feature points on them for decision-making. Moreover, when the static objects are getting close to the vehicle due to ego-motion, false alarms will be raised if their parallax across frames exceed the flow vector bound for the current frame. Besides, object mutual occlusion might also cause the false alarm occurrence when overlapping bounding boxes share the feature points for outlier rejection. Indeed, in these situations, the detection precision degrades. However, in practice, false alarms do not have fatal impact for the autonomous driving and misdetection of moving objects in the scene is not that critical when the objects are far away from the vehicle.

Table 2.1 – Comparison of MOS accuracy

Methods	Metrics	05	11	51	56
Our approach	Precision	**0.700**	**0.868**	**0.856**	**0.885**
	F-Score	0.762	**0.878**	0.773	**0.798**
Approach in [48]	Precision	0.690	0.696	0.680	0.768
	F-Score	**0.780**	0.792	**0.799**	0.777
Approach in [49]	Precision	0.383	0.675	0.556	0.510
	F-Score	0.513	0.770	0.706	0.664

2.3 Stereo-Vision based State Estimation and Uncertainty Modeling

The conventional monocular camera has the inherent problem of depth ambiguity for the state estimation and mapping tasks, see Fig. 2.15. It can be seen from Fig. 2.15 that, all the points which lie along the ray from the camera optical center **O** to the image point **p** uniformly satisfy the perspective projection rule. In order to distinguish the possible spatial locations of 3D point **P**, the stereo vision system is proposed in [1] for the 6 Degree of Freedom (DoF) motion estimation. The stereo vision system basically consists of two cameras that are set apart for the visual points triangulation (see Fig.

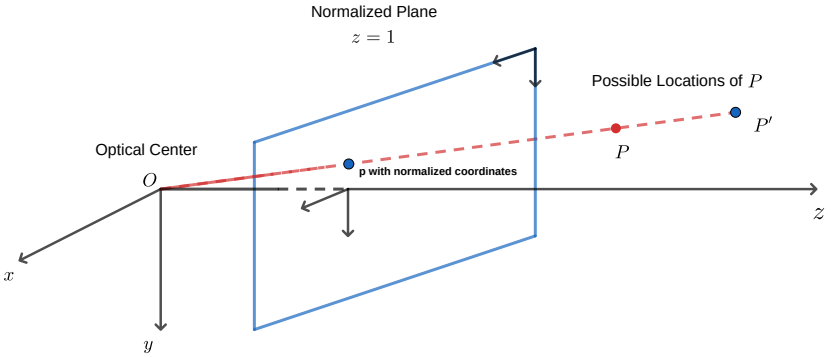


Figure 2.15 – Scheme of point depth ambiguity for the mono-vision system

2.16). This mechanism is similar to our human eyes, where the point depth can be easily

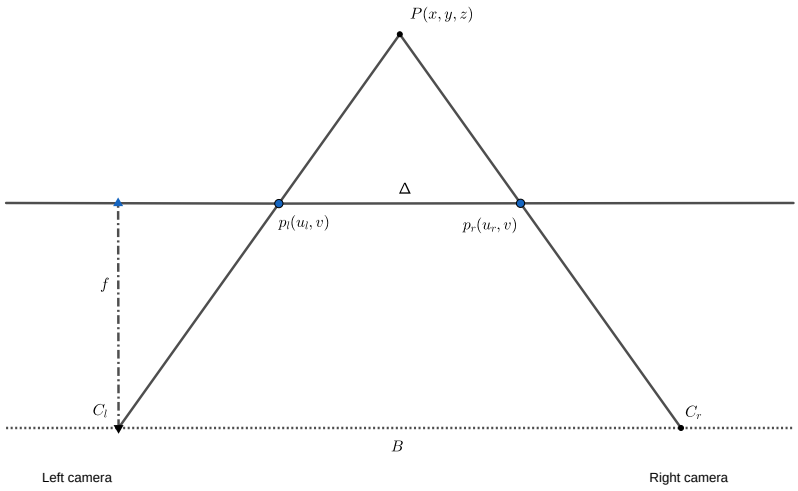


Figure 2.16 – Scheme of the stereo-vision system to estimate the depth of triangulated point P

deduced from the left-right pixel-wise disparity calculation. In Fig. 2.16, the 3D point $\mathbf{P}$ is captured by both the left and right cameras, where $\mathbf{p}_l$, $\mathbf{p}_r$ are the projected points on the image planes. Before the disparity and depth estimation, the captured left and right images need to be rectified to ensure the corresponding points stay on the same row, where the pixel v coordinates are identical. According to the principle of similarity of triangles, we can have equation as follows:

$$\frac{z-f}{z} = \frac{B-u_l+u_r}{B} \tag{2.15}$$

where z, f, B stand for the point depth, focal length and baseline distance respectively. After the simplification, we can obtain:

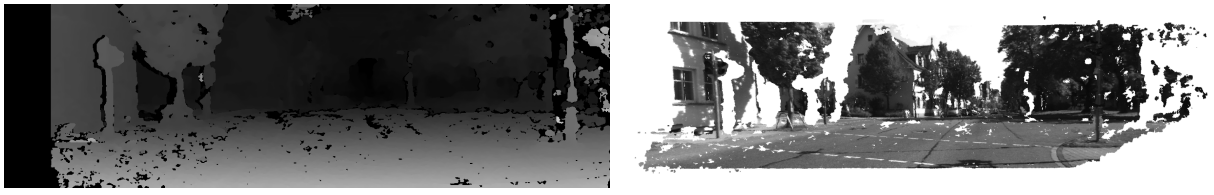
$$z = \frac{fB}{\Delta}, \quad \Delta \triangleq u_l - u_r \tag{2.16}$$

And Δ is defined as the disparity value between the corresponding left and right image points. As we can see from Eq. 2.16, the depth is inversely proportional to this disparity value. When the disparity value is obtained, to calculate the depth with Eq. 2.16 is an easy task. However, matching the exact points from left and right images is not that trivial for the stereo-vision system. Thus, the basic block matching method proposed in [57] solves the data association problem via comparing the Sum of Absolute Differences (SAD) of each block on the image plane. Furthermore, the Semi-Global Block Matching (SGBM) [58] approach additionally imposes the condition of similar disparity among neighboring blocks to improve the stereo matching accuracy, which makes the disparity calculation more reliable in complex scenes. It can be seen from Fig. 2.18 that the stereo image points

(a) The image captured by the left gray-scale camera in kitti dataset

(b) The image captured by the right gray-scale camera in kitti dataset

Figure 2.17 – The images captured with stereo-vision system in KITTI dataset



(a) The calculated disparity map with SGBM algorithm in KITTI dataset

(b) The reconstructed 3D scene map with the absolute depth information

Figure 2.18 – The disparity map and depth-aware scene reconstruction results in KITTI dataset

are assigned with the depth information for the scene reconstruction. Some of the parts

are left as blank due to the viewpoint changes or occlusion, which complicates the stereo matching and triangulation process. With the depth information, the 3D feature points can be directly leveraged for the 6 DoF ${}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}$ ego-motion estimation via minimizing the image plane re-projection errors, see Eq. 2.17.

$$ {}^{\mathcal{F}_{C_{t-1}}}\Theta^*_{\mathcal{F}_{C_t}} = \underset{{}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}}{\operatorname{argmin}} \sum_i \left\| {}^{\mathcal{F}_{I_t}}\mathbf{x}_i - {}^{\mathcal{F}_{I_t}}\hat{\mathbf{x}}_i \right\|^2_{\Sigma} \tag{2.17} $$

where $\| \cdot \|^2_{\Sigma}$ is the Mahalanobis distance with $\Sigma_{\mathbf{x}_i}$ as the covariance matrix for the i_{th} measurement and the uppercase $\mathcal{F}_{C_{t-1}}$, $\mathcal{F}_{C_t}$, $\mathcal{F}_{I_t}$ stand for the current camera frame, previous camera frame, current image frame respectively. The ${}^{\mathcal{F}_{I_t}}\hat{\mathbf{x}}_i = (\hat{u}_{i,l}, \hat{v}_{i,l}, \hat{u}_{i,r}, \hat{v}_{i,r})^T$ is the reprojected 2D point on the left and right images via inter-frame transformation ${}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}$ and camera perspective projection $\mathbf{K}$.

$$ {}^{\mathcal{F}_{I_t}}\hat{\mathbf{x}}_i = f(\mathbf{K}, {}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}, {}^{\mathcal{F}_{C_{t-1}}}\hat{\mathbf{P}}_{i-1}) \tag{2.18} $$

Since the 3D points are triangulated from the stereo images instead of adjacent monocular frames, the stereo-vision system will be more robust than the monocular scheme in challenging scenarios, such as aggressive motion and mutual occlusion [4].

The general formulation in Eq. 2.17 aims to find the relative transformation ${}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}$ that minimizes the image plane reprojection errors. The residual errors measure the distance from the estimation to measurement location on the image, which can be used to assess or quantify the estimation results without any priors. However, it needs to be noted that the data collected by the visual sensors are not noise-free, so the residual errors can not completely represent the estimation uncertainty. Furthermore, with more and more point matches taken into the residual optimization, the overall residual errors inevitably increase. In this case, residual errors alone are no longer suitable for the estimation uncertainty modeling. Intuitively, more point matches will consolidate the state estimation results and the estimation variance should decrease.

In order to quantify the optimum estimate uncertainty, we need to figure out how the estimation results are obtained. We assume that the measurement noise on the image plane observes the Gaussian distribution, with zero mean and the covariance $\Sigma_{\mathbf{x}_i}$. For the notation simplicity, the reprojection residual in Eq. 2.17 can be rewritten as:

$$ \|\mathbf{x} - f(\widehat{\boldsymbol{\Theta}})\|_{\Sigma} = \|(\mathbf{x} - \overline{\mathbf{x}}) - \mathbf{J}(\widehat{\boldsymbol{\Theta}} \ominus \overline{\boldsymbol{\Theta}})\|_{\Sigma} \tag{2.19} $$

where $\mathbf{x}$ represents the measurement value. $\widehat{\mathbf{\Theta}}$ is the estimated value and $\overline{\mathbf{x}}$, $\overline{\mathbf{\Theta}}$ stand for the true values. $\mathbf{J}$ is the Jacobian matrix which is the partial derivative $\frac{\partial f}{\partial \Theta}$ taken at the estimated $\widehat{\mathbf{\Theta}}$ and Σ denotes the measurement noise. The solution for the least square minimization follows the equation:

$$(\widehat{\mathbf{\Theta}} \ominus \overline{\mathbf{\Theta}}) = \left(\mathbf{J}^\top \Sigma^{-1}\ \mathbf{J}\right)^{-1}\ \mathbf{J}^\top \Sigma^{-1}(\mathbf{x} - \overline{\mathbf{x}}) \tag{2.20}$$

Then the estimated transformation can be expressed as:

$$\widehat{\mathbf{\Theta}} = \left(\mathbf{J}^\top \Sigma^{-1}\ \mathbf{J}\right)^{-1}\ \mathbf{J}^\top \Sigma^{-1}(\mathbf{x} - \overline{\mathbf{x}}) \oplus \overline{\mathbf{\Theta}} \tag{2.21}$$

If $\overline{\mathbf{x}}$ is contaminated with the Gaussian white noise with zero mean and known covariance matrix Σ, then the estimated parameters $\mathbf{\Theta}$ will also follow the Gaussian distribution, with the mean of $\overline{\mathbf{\Theta}}$ and covariance $\Sigma_{\mathbf{\Theta}}$

$$\Sigma_{\mathbf{\Theta}} = \left[\left(\mathbf{J}^\top \Sigma^{-1}\ \mathbf{J}\right)^{-1}\ \mathbf{J}^\top \Sigma^{-1}\right] \Sigma \left[\left(\mathbf{J}^\top \Sigma^{-1}\ \mathbf{J}\right)^{-1}\ \mathbf{J}^\top \Sigma^{-1}\right]^\top \tag{2.22}$$

After the simplification, we can obtain the backward covariance propagation equation:

$$\Sigma_{\mathbf{\Theta}} = \left(\mathbf{J}^\top \Sigma^{-1}\ \mathbf{J}\right)^{-1} \tag{2.23}$$

which propagates the uncertainty from the measurement space to the estimation space. The uncertainty backward propagation involves the first order linearization around the optimum estimate and we can concatenate all matched points factors to get the final transformation uncertainty:

$$\Sigma_{\Theta} = \mathbf{J}_{m2e}(\sum_i (\mathbf{J}_i^T(\Theta)\Sigma_{\mathbf{x}_i}^{-1}\mathbf{J}_i(\Theta)))^{-1}\mathbf{J}_{m2e}^T \tag{2.24}$$

where $\mathbf{J}_{m2e}$ is the Jacobian matrix to transform from the manifold estimation domain to the euclidean estimation domain, which will be detailed in Section 4.3. We can infer from Eq. 2.24 that, more inlier matches will aggregate the measurement information, which results a more confident estimation. And the quantitative results of the pose estimation uncertainty can be seen from Fig. 2.19, where the three sigma bounds can successfully bound the pose estimation errors.

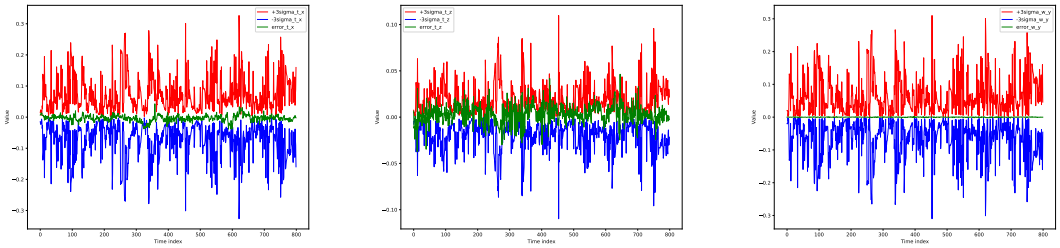


Figure 2.19 – Stereo-vision pose uncertainty along t_x, t_z, and r_y for KITTI sequence 03

2.4 Summary

In this chapter, both the pinhole monocular camera and stereo vision system are covered for the scene perception and self-localization. The camera relative transformation can be estimated by the reprojection residual errors minimization. Then concatenated relative poses need to be refined with the local bundle adjustment optimization, which can improve the visual odometry accuracy. However, the real-world environments are not always static. The moving objects such as pedestrians, cyclists, or vehicles may contaminate the data association process, which will result in biased ego-motion estimation and erroneous scene mapping. In order to segment the moving objects within the field of view, the geometric model and semantic information are adaptively combined to associate the semantic meaning to the geometric entities. The outlier rejection mechanism RANSAC is applied at a higher level abstraction to reject instance-level outliers which depends on the proportion of pixel-wise outliers in the bounding box. A moving object is extracted seamlessly from the semantic-guided RANSAC process and the computational complexity is reduced since sparse Shi-Tomasi corner features are used with semantic prior instead of all pixels in an image. Moreover, the fundamental matrix estimation process becomes more robust and efficient by taking the semantic prior into consideration. Without dedicated objects tracking and ego-motion estimation, our approach still achieves high precision and F-score on KITTI benchmarking sequences. Since the mono-vision perception has the inherent problem of depth ambiguity, the stereo-vision system is introduced to overcome this problem. The principle of the stereo-vision system is to capture synchronized images with both the left and right cameras, and depth of each matched pixel can be calculated with the left-right disparity and the baseline distance. After the outlier rejection, the relatively accurate 6 DoF estimation results can be obtained. Nonetheless, the deterministic ego-motion estimation scheme does not consider the visual sensor measurement

noise. In order to measure the estimation uncertainty, the first-order linear approximation is made to calculate the Jacobian matrix. The first-order Jacobian matrix helps to propagate the uncertainty from the measurement space to the estimation space. As a result, the confidence of the obtained optimum estimate can be assessed, which is vital for the conservative and secure decision-making.

Even though the visual cameras are versatile for the geometric and semantic parsing of the scene, they tend to have poor performance in low illumination condition. Thus, the full autonomous perception system tends to fuse the visual sensors with other sensing modalities, to increase the system resilience and robustness against measurement noise and perception degeneration cases. In Chapter 4, the LiDAR-Vision sensor fusion approach is developed for more robust perception and ego-motion estimation.

Chapter 3

LiDAR-based Perception and State Estimation

The previous Chapter focuses on leveraging the visual sensors for the moving objects segmentation and state estimation. The state estimation is two-fold, which includes the ego-motion estimation and map points registration. Indeed, in order to navigate in unknown real-world environments, accurate self-positioning and precise perception of the surroundings are two fundamental capabilities for intelligent vehicles. With the advancement of the automobile industry and mobile robotics, different sensing modalities are employed onboard to analyze the scene and to estimate the vehicle ego-pose. For outdoor maneuvering, Global Navigation Satellite Systems (GNSS) are widely used solutions to obtain the vehicle pose when a reliable signal can be obtained from several satellites. However, in GPS-denied areas, the popular high-definition Light Detection and Ranging (LiDAR) sensor is gaining popularity due to its inherent high precision and robustness against noise. Compared to visual sensors, LiDAR is invariant to the illumination conditions and its omnidirectional field of view enables robust scale-aware pose estimation and precise 3D scene mapping.

Basically, the multi-layer LiDAR measurements are consisted of dense point clouds that are collected from the laser beam reflection. Given the consecutive timestamped 3D LiDAR scans, a rigid transformation can be obtained to tightly align the associated LiDAR point clouds, which forms the basis for LiDAR-based ego-motion estimation. By matching the corresponding LiDAR points, the LiDAR-based ego-motion estimation aims to determine the position and orientation of the ego-vehicle in an incremental way and the absolute pose to the reference frame can be retrieved by means of rigid transformation accumulation. Moreover, the point clouds returned from the laser reflection provide the reflectance value and rich geometric features that enable real-time 3D object detection. For autonomous vehicles, it is essential to predict the location and class of the traffic participants such as pedestrians, cyclists and cars in the neighboring areas. Due to the

3D perception capability of the LiDAR sensor, the objects are usually recognized and described with oriented 3D bounding boxes in the ego-vehicle sensor frame. The 3D bounding boxes give the center coordinates and the width, length and height dimensions of the detected objects, where the orientation angle is simplified as the yaw angle with respect to the ego-vehicle sensor frame.

3.1 3D Semi-Direct LiDAR Scan Matching

In the past few decades, several achievements in pose estimation using 3D LiDAR scan matching-based approaches have been obtained, which reached centimeter-level precision. In order to tightly align two time-consecutive point clouds, ICP-based algorithms are well-recognized methods to estimate the relative pose of the LiDAR frames. The direct ICP scan matching algorithm was firstly introduced in [59]. The ICP algorithm iteratively searches for the nearest neighbor in the target point cloud and builds the correspondences, based on an objective function, that refines the optimal relative transformation through iterations. In [59], the point-to-point distance is used for the closest neighbor association, which might be too greedy when the two scans are captured with strong viewpoints changes. In order to mitigate this problem, the point-to-plane distance is proposed in [60] for robust data association. Compared with the point-to-point distance metric, the point-to-plane distance metric relaxes the strict point-to-point correspondence restrictions, which is more adaptable for partially overlapped scans in practice. In order to speed up the scan matching data association process, a KD-tree data structure [61] is adopted for efficient nearest neighbor queries. To increase the robustness against outliers, the Trimmed ICP algorithm is presented in [62]. In this work [62], the strict selection of low mean square error correspondences allows a scan matching with low overlaps. With the purpose of unifying the point-to-point and point-to-plane distance metrics into a probabilistic framework, Generalized-ICP (G-ICP) has been proposed in [63]. In the G-ICP framework, the scan alignment accuracy, efficiency and robustness are maintained under a variety of challenging situations. Instead of aligning the whole raw 3D point cloud, the feature-based methods extract distinctive feature points and build the correspondences based on high-dimension descriptors instead of simple Euclidean distance between the points positions. For example, a double-layer feature-based method [64] extracts the ground and vertical features to represent the scene, and robustly associates the correspondences between the scan observations. Moreover, the edge and planar points

are classified and associated according to the curvature value in LOAM [15] to realize the long-term localization. In order to further alleviate the outliers contamination, T-LOAM [65] pre-classified the points into edge features, sphere features, planar features and ground features, which are associated with a truncated least-squares method to maintain the scan matching robustness and accuracy at the same time. For large-scale scene localization, the multi-scale feature maps are built as a prior in [66] for efficient global positioning. Recently, inspired by images-based approaches, more and more viewpoint invariant descriptors such as PFH [67], FPFH [68], and SHOT [69] emerged to characterize the local patches around the points of interest in the LiDAR scans. Sample Consensus Initial Alignment (SAC-IA) has been proposed in [68] for the rough scan alignment. This approach is robust to sensor noise but might cause overfitting in information-deprived environments such as indoor corridors, outdoor tunnels or highways. The accuracy of the standard ICP scan matching is highly dependent on the initialization process as reported in [70]. It means that a large deviation of the initial alignment may cause the ICP optimization divergence or being stuck in local minima. Thus, hybrid methods [21] [71] [72] [73] [74] [19] that incorporated visual and range sensor information have been proposed for reliable state estimation and scene reconstruction. In [21], the omnidirectional visual features (assigned with the SIFT [30] descriptors) are back-projected to the LiDAR sensor frame to boost the ICP convergence. The visual and LiDAR sensor information are loosely coupled by covariance intersection in [71] for the robust inter-frame pose estimation. Besides, a LiDAR-Camera joint optimization approach has been proposed in [72] for accurate pose estimation and dense 3D model reconstruction. At the same time, the sparse LiDAR depth map can also be enhanced with the involvement of the state-of-the-art event-based camera to realize dense scene reconstruction [73]. Moreover, a polarized camera factor can be integrated in a factor graph framework [74] to correct the yaw angle estimation in sparse and GPS-denied areas. Furthermore, the drift of the incremental LiDAR scan alignment can be eliminated by the vision-based place recognition technique, as demonstrated in [19]. With the advent of deep learning era, data-driven local 3D descriptors such as 3DMatch [75], PPFNet [76], and 3DFeat-Net [77] appeared, which learned from pre-built correspondences in the training dataset and encoded the volumetric patches into descriptive vectors for robust feature matching. Besides, an Long Short-Term Memory (LSTM) [78] network was put forward for real-time LiDAR odometry estimation, which inferred the possible connections across consecutive frames. Another deep neural network named 3DRegNet was then proposed in [79] for robust correspondences outliers rejection

and efficient 3D point cloud alignment with noisy input. Moreover, a modified siamese network OverlapNet [80] has achieved prosperous results by adaptively comparing the similarity between pairs of LiDAR scans to address the challenges of purely LiDAR-based loop closure and global localization.

Depending on whether raw data and local features are extracted and matched in the source and target point clouds, LiDAR scan matching algorithms can be roughly categorized into two branches, namely the direct and feature-based point cloud alignment approaches. On the one hand, the direct scan matching method aligns the raw point clouds without distinguishing the feature points. The correspondences are built iteratively according to the nearest neighbor criteria, and the optimal transformation is estimated as the relative pose between the source and target point clouds. However, it needs to be noticed that the direct method always requires a good initial guess to start, and it does not work efficiently for small overlapping and noisy scans. On the other hand, the feature-based scan matching methods extract local features from raw point clouds and build high dimension descriptors for the features' matching process. The established corresponding points will then be used to estimate the relative pose, while increasing the robustness against noise and erroneous matches. Nonetheless, in repetitive and feature-less environments, the performance of the feature-based method may degrade due to the lack of features working as anchors during the alignment process. Semi-direct approaches were firstly proposed in [81] [82] [83] for vision-based state estimation to overcome the problem of visual illumination changes and loss of visual feature tracking. Basically, the estimated transformation from the feature-based method serves as the starting point for the direct method to increase the estimation precision and robustness. Inspired by those performances in vision-based state estimation, a so-called semi-direct scan matching approach is proposed that combines the conceptually complementary direct and feature-based scan matching. The proposed semi-direct scan matching algorithm ensures that the registration is robust against high-speed or large-rotation motion, and travel in repetitive and feature-less environments. Besides, for vehicles driving on roads, the ground points that appear to be highly isotropic, are inevitably scanned and matched, which may bias the state estimation. In order to reduce the registration lag effects, ground points are removed beforehand. Since the ground vehicle platform is used for scene perception and state estimation, the roll and pitch angles will be very small and are mainly due to the vehicle body vibration. So they can be safely neglected and will not cause critical issues. Thus, a simple but efficient method is applied to identify and to clear the ground points

with the prior vehicle height and planar motion assumptions.

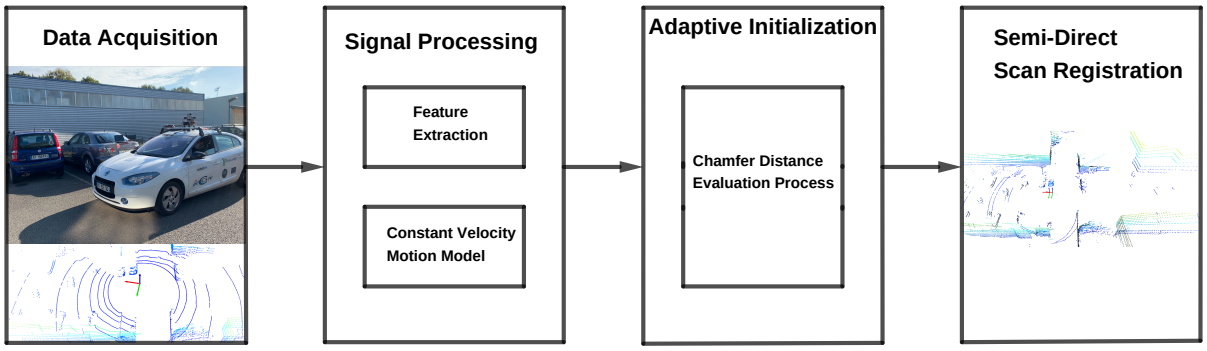


Figure 3.1 – Adaptive Semi-Direct LiDAR Scan Matching for Ground Vehicles Localization and Mapping.

The pipeline of the proposed semi-direct LiDAR scan matching approach is shown in Fig. 3.1. For consecutive LiDAR scans perceived in the local frame, point cloud pre-processing is necessary before implementing the scan matching. The point cloud pre-processing includes scattered outliers and ground points removal, which will significantly ease the false correspondence matching problem. Afterwards, a robust semi-direct scan matching that adopts the adaptive initialization strategy is applied to compute the relative poses.

The scattered scene components such as tree leaves or bushes are sparse and non-permanent, which may cause a challenge for accurate scan matching. Besides that, due to the existence of measurement noise, some spurious points may be collected and matched which may corrupt the alignment results. It is thus mandatory to filter out such scattered outliers in the point clouds before implementing the scan alignment. In the proposed approach, a Gaussian distribution based outlier rejection method is performed to remove the sparse and isolated points. Firstly, for each 3D point $\{\mathbf{x}_i\}$, the distances to its neighbors $\mathbf{x}_j \in \mathbf{N}(\mathbf{x}_i)$ are computed within a pre-defined area. In this thesis, the query area $\mathbf{N}(\mathbf{x}_i)$ is defined as the 20 nearest neighbors around the point $\{\mathbf{x}_i\}$. With the assumption that the calculated distances satisfy a Gaussian distribution, the mean $\mu(\mathbf{x}_i)$ and standard deviation $\sigma(\mathbf{x}_i)$ are computed. Then, the neighbors $\{\mathbf{x}_j\}$ with distances exceeding the sigma-rule boundary from the mean distance, are marked as outliers to be trimmed. Subsequently, the voxel grid filter with the leaf size L_{size} is applied to down-sample and to approximate the point clouds for efficient scan alignment. For each voxel $\mathbf{V}_i$, the points $\{\mathbf{x}_i\}$ inside it, are approximated with their centroid $\mathbf{C}_i = \sum_{\mathbf{x}_i \in \mathbf{V}_i} \frac{\mathbf{x}_i}{|\mathbf{V}_i|}$.

This operation will greatly reduce the number of points without losing the raw points distribution. When traveling on real roads, the point cloud acquired by the LiDAR sensor placed on the top of the vehicle naturally contains many ground points. Ground points on flat roads tend to be isotropic and they encode little geometric information for the data association step. During the iterative nearest neighbor tracking, a slight deviation may cause mismatches of the ground points and distort the whole scan alignment. This usually occurs in city or suburb open areas where the navigable space is the predominant planar surface. Therefore, ground points removal is crucial to ease the scan matching process. To address these problems, a consensus-based method is applied to segment the ground points with the assumption that the vehicle follows a planar motion and the LiDAR vertical position is known. On this basis, a set of potential ground points $\{\hat{\mathbf{G}}_t : \mathbf{x}_l^G \in \hat{\mathbf{G}}_t\}$ are selected, incorporating all the points that locate h_{prior} along the z-axis below the vehicle roof. Here, h_{prior} is highly linked to the LiDAR installation height. Meanwhile, another condition for potential points selection is that the unit normal vector $\vec{\mathbf{n}}_{\mathbf{x}_l}^G$ around $\mathbf{x}_l^G$ needs to stay within an offset of γ_{off} from the z-axis of the LiDAR frame. The offset γ_{off} allows slightly sloped terrain points to participate in the ground plane estimation, considering the fact that the driving road is not always fully flat.

$$\mathbf{x}_{lz}^G < -h_{prior}, \quad \vec{\mathbf{n}}_{\mathbf{x}_l}^G \cdot [0,0,1]^T > cos(\gamma_{off}) \tag{3.1}$$

After that, the consensus-based method iteratively picks three non collinear points $\{\mathbf{x}_i^G, \mathbf{x}_j^G, \mathbf{x}_k^G\}$ within the pre-defined point set $\{\hat{\mathbf{G}}_t\}$ to fit the ground plane equation $\{\hat{\mathbf{P}}_G : Ax + By + Cz + D = 0\}$. The distances of the potential ground points $\{\mathbf{x}_l^G \in \hat{\mathbf{G}}_t\}$ to the estimated plane $\hat{\mathbf{P}}_G$ are then computed and summed up to vote for the plane candidates until the convergence criteria are met. At the end of the iterations, the plane with the least point-to-plane distances is chosen as the ground plane $\mathbf{P}_G^*$. As long as the ground plane is found, the points in $\{\hat{\mathbf{G}}_t\}$ with the distances less than h_{off} to the ground plane $\mathbf{P}_G^*$ are also considered as belonging to the ground to increase robustness against measurement noise (see Fig. 3.2). In this figure, the estimated ground plane is $\{\mathbf{P}_G^* : -0.00x + 0.03y + 1.00z + 1.75 = 0\}$, which reaches centimeter level precision according to installation measurements. Finally, the pre-processing procedure ends, and the ground points excluded from LiDAR scans $\{\mathbf{S}_t^{\triangle}\}$ are used to minimize the alignment error.

As reported in [70], the direct scan matching with raw point clouds tends to be less

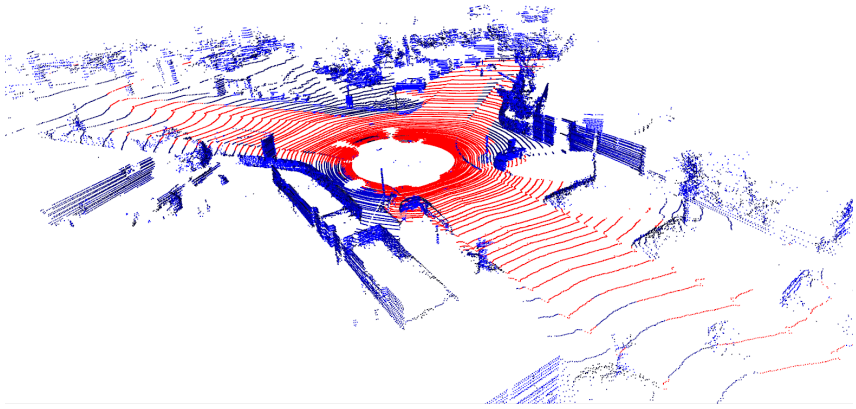


Figure 3.2 – Visualization of the ground points (in red) segmented by our consensus-based method with the KITTI HDL-64E LiDAR

efficient with large baseline motions, while feature-based methods lose their advantages of bootstrapping from local minima in feature-less environments. To complement the weakness of individual direct or feature-based methods, a hybrid semi-direct approach is proposed, to allow robust estimation under challenging conditions. Here, a feature-based method is leveraged to estimate a coarse but globally consistent inter-frame pose that serves as a prior for the following multi-scale direct dense point clouds alignment. In this proposed semi-direct approach, the point-to-plane distance metric is adopted for the optimal estimation of $\mathbf{T}^*$,

$$\mathbf{T}^* = arg\min_{T} \sum_{(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}} \|(\mathbf{x}_i^{'} - \mathbf{T}\mathbf{x}_i)\mathbf{n}_i^{'}\|^2 \tag{3.2}$$

where $\mathbf{C}_{ii'}$ is the correspondence set incorporating the point pairs in the source and target LiDAR scans. The plane normal vectors $\mathbf{n}_i^{'}$ around $\mathbf{x}_i^{'}$ in the target LiDAR scan could effectively guide the scan matching process to distinguish points lying on different surfaces and discard unreliable correspondences $\{(\mathbf{x}_i, \mathbf{x}_i^{'})\}$. For the sake of reducing spurious correspondences, the selected features for the sparse point cloud alignment need to be distinctive and invariant to viewpoint changes. To this end, intrinsic Shape Signatures (ISS) [84] is a salience-based keypoints extraction method, in which the salience measure is derived from the scatter matrix $\mathbf{\Sigma}(\mathbf{x}_i)$ eigenvalue decomposition.

$$\bar{\mathbf{x}} = \frac{1}{|\mathbf{S}(\mathbf{x_i})|} \sum_{\mathbf{x}_j \in \mathbf{S}(\mathbf{x}_i)} \mathbf{x}_j \tag{3.3}$$

$$\mathbf{\Sigma}(\mathbf{x}_i) = \frac{1}{|\mathbf{S}(\mathbf{x_i})|} \sum_{\mathbf{x}_j \in \mathbf{S}(\mathbf{x}_i)} (\mathbf{x}_j - \bar{\mathbf{x}}) (\mathbf{x}_j - \bar{\mathbf{x}})^T \tag{3.4}$$

where $\mathbf{S}(\mathbf{x}_i)$ are the spherical neighbours of $\mathbf{x}_i$ within a pre-defined radius. With the scatter matrix $\mathbf{\Sigma}(\mathbf{x}_i)$ eigenvalue decomposition, the descending eigenvalues $\{\lambda_i^1, \lambda_i^2, \lambda_i^3\}$ can be obtained, and their corresponding eigenvectors $\{e_i^1, e_i^2, e_i^3\}$ are used to build the intrinsic reference frame. To ensure the distinctiveness of the local features, ratios of the sequential eigenvalues are required to not exceed the thresholds γ_{21} and γ_{32}.

$$\lambda_2^2/\lambda_1^2 < \gamma_{21}, \quad \lambda_3^2/\lambda_2^2 < \gamma_{32} \tag{3.5}$$

The thresholds γ_{21} and γ_{32} guarantee that the established intrinsic reference frame exhibits maximum salience along the principal directions, which makes features more informative and recognizable from various viewpoints. As long as the ISS features are extracted from the 3D point cloud, the efficient Fast Point Feature Histogram (FPFH) descriptor [68] is used for robust data association. FPFH is basically a 33-dimensional vector that characterize the local geometry around a point. It efficiently increases the keypoints description and helps to establish the keypoints correspondences. Through iteratively taking three pairs of matched feature points $\{(\mathbf{x}_i, \mathbf{x}_i^{'}), (\mathbf{x}_j, \mathbf{x}_j^{'}), (\mathbf{x}_k, \mathbf{x}_k^{'})\}$ and implementing correspondence consistency (edge similarity) check, the transformation matrix $\mathbf{T}_f$ that optimally aligns the sparse feature points could be obtained by using the Random Sample Consensus (RANSAC) method. The correspondence consistency check is to verify the edges distance $\{(d_{ij}, d_{ik}, d_{jk}), (d_{ij}^{'}, d_{ik}^{'}, d_{jk}^{'})\}$ formed by the features in each frame. This prevents mismatches in the environment with repeatable features. The correspondences are then considered as valid, if the features are not collinear

$$\begin{aligned} (\mathbf{x}_i - \mathbf{x}_j)/d_{ij} &\neq (\mathbf{x}_i - \mathbf{x}_k)/d_{ik} \\ (\mathbf{x}_i^{'} - \mathbf{x}_j^{'})/d_{ij}^{'} &\neq (\mathbf{x}_i^{'} - \mathbf{x}_k^{'})/d_{ik}^{'} \end{aligned} \tag{3.6}$$

and their formed edges have similar length as

$$\begin{aligned} 0.9 < d_{ij}/d_{ij}^{'} < 1.1 \\ 0.9 < d_{ik}/d_{ik}^{'} < 1.1 \\ 0.9 < d_{jk}/d_{jk}^{'} < 1.1 \end{aligned} \tag{3.7}$$

Given a set of associated feature points, the RANSAC-based scan matching does not require initial guess and is more robust to outliers. This alleviates the problem of getting

stuck at the local minima. However, in the feature-less environment with few distinctive features to be extracted, feature tracking tends to be hard to manage. In this way, we need to evaluate the estimated pose $\mathbf{T}_f$ from the feature-based method and the last pose estimation heuristics ${}^t\hat{\mathbf{T}}_{t-1}$ with the Chamfer Distance (CD), to choose a more reliable initialization point $\mathbf{T}_{init}$ for the dense alignment.

$$\mathbf{T}_{init} = arg \min_{{}^t\hat{\mathbf{T}}_{t-1}, \mathbf{T}_f} \{CD({}^t\hat{\mathbf{T}}_{t-1}\mathbf{S}_t^{\triangle}, \mathbf{S}_{t+1}^{\triangle}),\ CD(\mathbf{T}_f\mathbf{S}_t^{\triangle}, \mathbf{S}_{t+1}^{\triangle})\} \tag{3.8}$$

where the notation $(\cdot)^{\triangle}$ stands for ground points free LiDAR scans, and the Chamfer Distance (CD) is a metric to measure the tightness of two aligned point clouds, with the expression as:

$$\begin{aligned} \mathrm{CD}\left(\mathbf{S}_1, \mathbf{S}_2\right) = & \frac{1}{|\mathbf{S}_1|} \sum_{\mathbf{x}_i \in \mathbf{S}_1} \min_{\mathbf{x}_i^{'} \in \mathbf{S}_2} \|\mathbf{x}_i - \mathbf{x}_i^{'}\|^2 \\ & + \frac{1}{|\mathbf{S}_2|} \sum_{\mathbf{x}_j^{'} \in \mathbf{S}_2} \min_{\mathbf{x}_j^{\in} \mathbf{S}_1} \|\mathbf{x}_j - \mathbf{x}_j^{'}\|^2 \end{aligned} \tag{3.9}$$

Then the multi-scale pyramidal [85] dense point cloud alignment is implemented to refine the coarse initial pose $\mathbf{T}_{init}$. The multi-scale pyramid defines a two-layer maximum correspondence distance as $3 \times L_{size}$ and L_{size}, and L_{size} is the downsampling voxel leaf size. The first layer point cloud alignment has the maximum correspondence distance of $3 \times L_{size}$, whose convergence criteria is simple to reach. As a result, the first layer alignment further eliminates the effect of false correspondence and reduces the risk of being stuck in the local minima. Then, based on the result from the first layer, the second layer pyramid searches in a finer scale for the final tight point-cloud alignment. This design adaptively determines the nearest neighbor searching radius without fine-tuning, which greatly facilitates optimization convergence for partially overlapped point clouds. Since more raw points information is considered, the point cloud alignment is more robust in feature-less scenes with few distinctive 3D feature points to track.

To evaluate the performances of the proposed Semi-Direct ICP (SD-ICP) scan matching approach, extensive experiments have been carried out using the public KITTI [55] and self-recorded LS2N datasets at the Centrale Nantes Campus. The KITTI dataset point cloud is acquired with a Velodyne HDL-64E laser scanner, which has 64 channels with a maximum range of 120 m. Our self-recorded dataset is collected with a light-weight Velodyne VLP-16 LiDAR, which only has 16 perception channels and renders sparse point clouds. The parameters chosen for the proposed approach are summarized in the Tab.

Algorithm 2 Semi-Direct Scan Matching Algorithm

Input: Consecutive LiDAR scans $\{\mathbf{S}_t, \mathbf{S}_{t+1}\}$, Relative transformation heuristics ${}^{t}\hat{\mathbf{T}}_{t-1}$
Output: Estimated transformation ${}^{t+1}\hat{\mathbf{T}}_{t}$ that tightly aligns the consecutive scans

1: Initialize ${}^{1}\hat{\mathbf{T}}_{0} \leftarrow \mathbf{I}_4$
2: **while** New LiDAR scan arrives **do**
3: Remove the scattered outliers and ground points to get the processed LiDAR scans $\{\mathbf{S}_t^{\triangle}, \mathbf{S}_{t+1}^{\triangle}\}$
4: Apply the ISS keypoints detection, FPFH description to $\{\mathbf{S}_t^{\triangle}, \mathbf{S}_{t+1}^{\triangle}\}$
5: Implement the keypoints correspondence consistency check and only keep the valid ones using Eq. 3.6, 3.7
6: Obtain the rough transformation $\mathbf{T}_f$ with the consistent keypoints correspondences using RANSAC
7: Evaluate the Chamfer Distance (CD) of ${}^{t}\hat{\mathbf{T}}_{t-1}$ and $\mathbf{T}_f$, then choose the initial guess $\mathbf{T}_{init}$ using Eq. 3.8
8: Conduct the multi-scale dense alignment with $\mathbf{T}_{init}$ and get the relative transformation ${}^{t+1}\hat{\mathbf{T}}_{t}$
9: Update the LiDAR scan timestamp $t \leftarrow t + 1$
10: **end while**

3.1. The considered evaluation metrics are the Relative Fitness (RF) and Relative Root

Table 3.1 – The parameters table for the proposed SD-ICP

L_{size}	h_{prior}	γ_{off}	h_{off}	γ_{21}	γ_{32}
0.1 m	1.4 m	$\pi/5$ rad	0.2 m	0.975	0.975

Mean Square Error (RMSE) of the inlier correspondences $\{\mathbf{C}_{ii'}\}$, that can be expressed as:

$$RF = \frac{|\mathbf{C}_{ii'}|}{|\mathbf{S}_{t+1}|}, \quad RMSE = \frac{1}{|\mathbf{C}_{ii'}|} \sum_{(\mathbf{x}_i, \mathbf{x}_i') \in \mathbf{C}_{ii'}} \sqrt{\|\mathbf{x}_i - \mathbf{x}_i'\|^2} \tag{3.10}$$

On the one hand, the Relative Fitness (RF) measures the proportion of associated inliers $\{\mathbf{C}_{ii'}\}$ among the full target cloud $\{\mathbf{S}_{t+1}\}$, and higher relative fitness values represent better scan matching results. On the other hand, RMSE measures the root mean square errors of all inlier correspondences, and lower RMSE stands for tighter alignment.

According to the KITTI sensor setup datasheet [1], the LiDAR installation height is 1.73 m from the ground, which verifies that our ground plane estimation in Fig. 3.2 is precise and reliable. This can be mainly attributed to the fact that the potential ground

1. http://www.cvlibs.net/datasets/kitti/setup.php

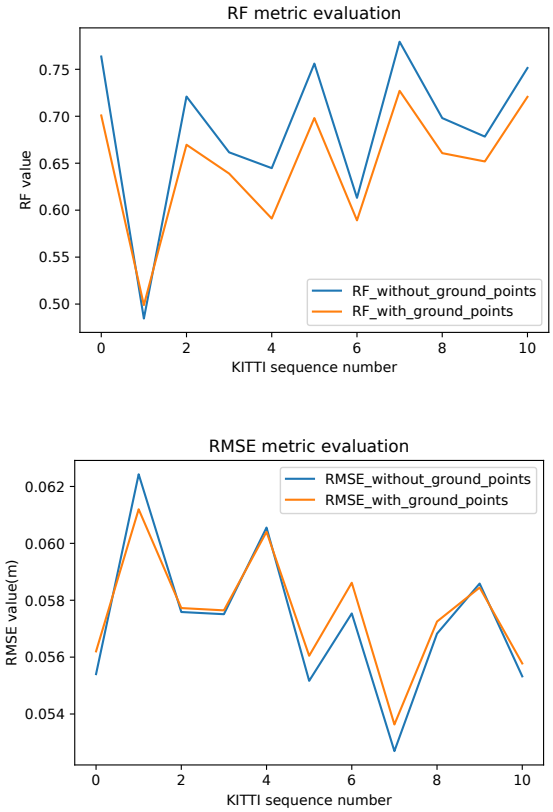
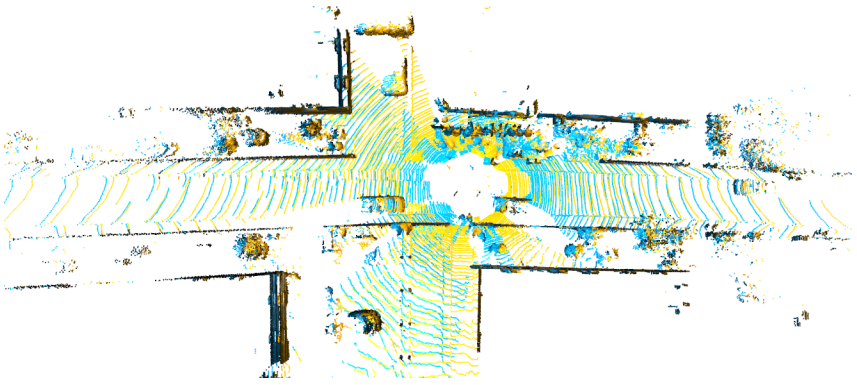


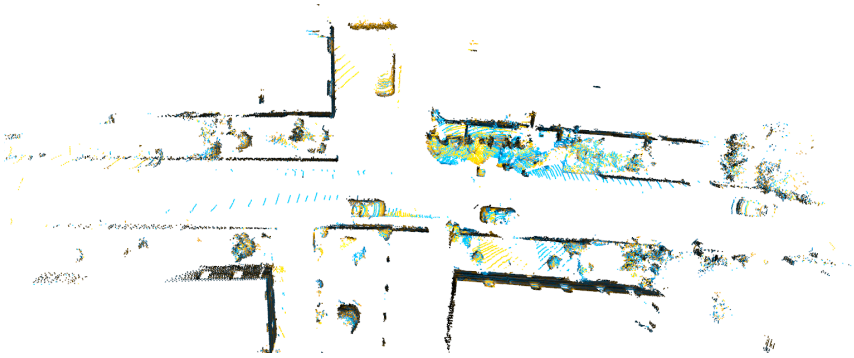
Figure 3.3 – Quantitative relative fitness and RMSE metrics evaluation of scan matching with(out) ground points for KITTI odometry sequence 00-10

points set has been strictly selected and only contains a few outliers. As a result, the consensus-based ground plane estimation avoids local minima, which facilitates accurate ground points segmentation. The registration lag effect, which means that the ground points erroneous matches distract the point clouds from seamless alignment, is shown in Fig. 3.3. It can be seen that ground points' existence lowers the inliers proportion and amplifies the relative RMSE value, which reveals the necessity to remove the ground points before conducting scan matching. However, for KITTI sequence 01, the ego-vehicle travels on the highway and is constantly surrounded by other moving vehicles, which poses great challenges for accurate ego-pose estimation. In this case, removing the ground points does not improve the scan matching performance. To integrate an Inertial Measurement Unit for ego-pose pre-integration may ease this problem.

In order to benchmark the performances of the proposed semi-direct scan matching method, a thorough quantitative evaluation is conducted in various scenarios such as ur-



(a) The ground points erroneous matches distract the point clouds from seamless alignment



(b) The registration lag effect is greatly reduced by ground points removal

Figure 3.4 – Qualitative results for the registration lag effect caused by the ground points erroneous matches for the KITTI sequence 05

ban (KITTI 04), highway (KITTI 01), residential (KITTI 00, 02-03, 05-10) and outdoor parking (LS2N 00-01) areas. In these different driving scenarios, the vehicle ego-motion varies a lot, which provides high speed motion in the highway and mild motion in the parking area for us to investigate the performance of different LiDAR scan matching approaches. Tab 3.2 and Tab 3.3 [2] list the relative fitness and RMSE metric values of the state-of-the-art scan matching methods as well as our approach, including Direct ICP [60] (D-ICP), Feature-based ICP [68] (F-ICP), Generalized ICP [63] (G-ICP) and our Semi-Direct ICP (SD-ICP). From Tab 3.2 and Tab 3.3, it is shown that the dense point cloud alignment using the direct ICP is generally superior to the sparse feature-based ICP. Particularly, for the sequence of LS2N 00, direct ICP outperforms other approaches in the term of the fitness metric due to the fact that the ego-vehicle moves slowly in the parking

2. The scan registration with the relative fitness below 30% is considered as invalid and the RMSE value is not calculated in that case

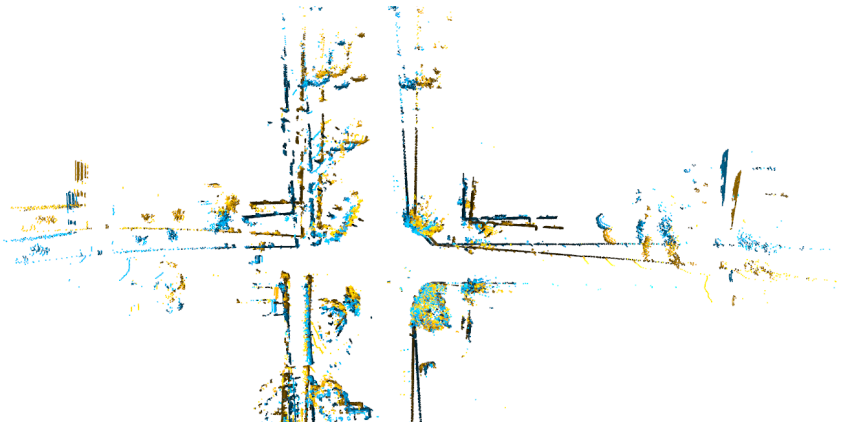
Table 3.2 – SD-ICP Registration Results Benchmarking with RF(%)

Metric / Dataset	D-ICP	F-ICP	G-ICP	SD-ICP
KITTI_00	47.7%	27.0%	59.7%	**77.2%**
KITTI_01	26.8%	24.8%	27.2%	**49.4%**
KITTI_02	38.1%	29.9%	57.7%	**73.2%**
KITTI_03	47.2%	27.7%	62.8%	**67.2%**
KITTI_04	31.7%	30.9%	31.6%	**65.4%**
KITTI_05	43.8%	28.9%	61.4%	**76.4%**
KITTI_06	27.6%	19.8%	40.9%	**62.3%**
KITTI_07	56.8%	33.4%	62.4%	**78.7%**
KITTI_08	39.1%	24.5%	50.4%	**70.7%**
KITTI_09	28.3%	21.4%	52.6%	**68.8%**
KITTI_10	49.5%	30.7%	69.8%	**76.0%**
LS2N_00	**74.7%**	69.9%	74.3%	72.7%
LS2N_01	44.8%	32.9%	44.9%	**49.5%**
Average	42.8%	30.9%	53.5%	**68.3%**

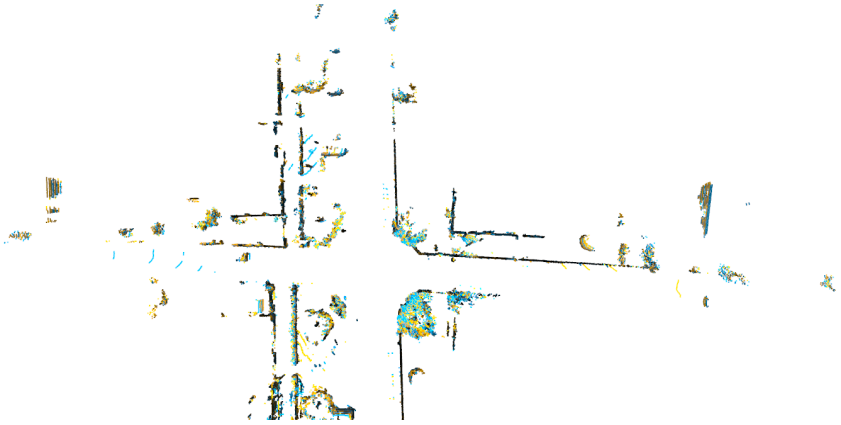
Table 3.3 – SD-ICP Registration Results Benchmarking with RMSE (CM)

Metric / Dataset	D-ICP	F-ICP	G-ICP	SD-ICP
KITTI_00	5.83	–	5.68	**5.53**
KITTI_01	–	–	–	**6.23**
KITTI_02	6.18	–	5.94	**5.75**
KITTI_03	6.08	–	5.83	**5.75**
KITTI_04	**5.91**	6.03	5.92	6.05
KITTI_05	5.80	–	5.64	**5.51**
KITTI_06	–	–	5.84	**5.75**
KITTI_07	5.49	5.98	5.36	**5.26**
KITTI_08	6.05	–	5.91	**5.68**
KITTI_09	–	–	6.07	**5.85**
KITTI_10	5.95	6.38	5.62	**5.53**
LS2N_00	5.37	5.56	5.41	**5.34**
LS2N_01	6.11	6.40	6.07	**6.02**
Average	5.88	6.07	5.77	**5.71**

area. Since more raw point clouds information is leveraged for scan matching, the direct ICP has better performance when the subsequent scans that share adequate overlapping areas. The sparsity of the point clouds obtained by 16-layer LiDAR is another reason



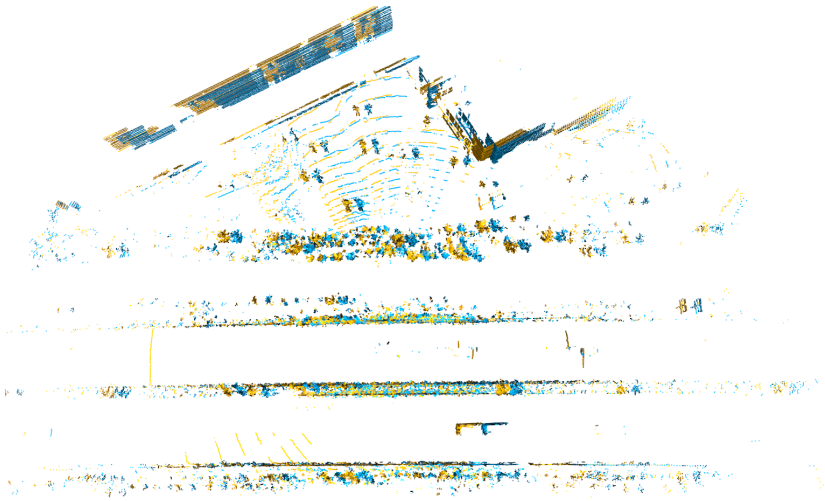
(a) Point cloud registration stuck in local minima for the D-ICP



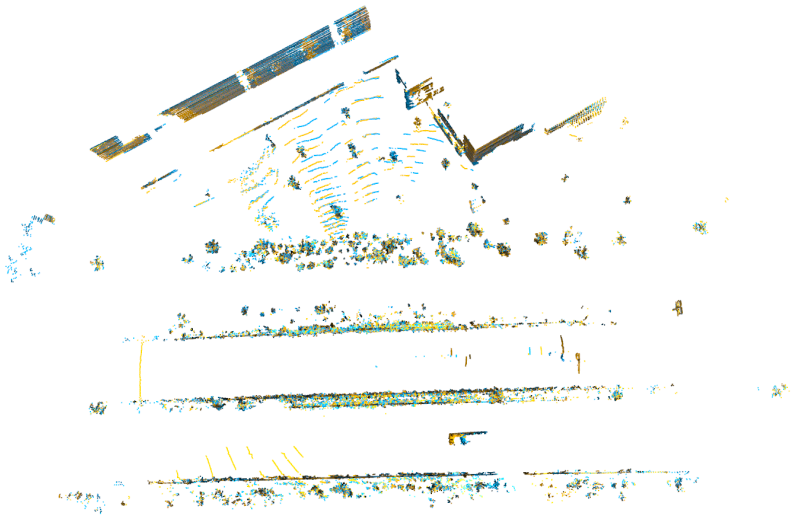
(b) Point cloud registration boosted from local minima for the SD-ICP

Figure 3.5 – Point cloud registration results visualization at the road intersection for frame 133-134 in the KITTI sequence 05

for the poor performance of the feature-based method. However, when the ego-vehicle experiences important viewpoint changes, the direct ICP loses its advantages, especially with the poor identity matrix transformation initialization. Due to the unanticipated decrease of overlaps, the inherent iterative nearest neighbor association strategy of D-ICP is vulnerable and tends to be stuck in the local minima, see Fig. 3.5(a). On the contrary, since high dimensional feature descriptors are invariant to viewpoint changes, the feature-based scan matching is more robust for the large baseline motion across consecutive frames. The putative descriptor-based keypoints correspondences are constructed beyond iterative nearest neighbor searching and update loop, which reduces the risks of being stuck in the local minima. With pre-established correspondences, the initialization-free sampling consensus-based method is applied to reject potential outliers and to obtain



(a) Point cloud registration lag in the feature-less environment for the F-ICP



(b) Robust point cloud registration with the adaptive initialization strategy for the SD-ICP

Figure 3.6 – Point cloud registration results visualization on the highway for frame 634-635 in the KITTI sequence 01

the optimal relative transformation. Nevertheless, it also needs to be mentioned that the feature-based scan matching is highly dependent on the feature detection procedure, which may lead to inaccurate or biased registration in the feature-less or feature-repetitive environments such as corridors or highways, see Fig. 3.6(a). The drawbacks of the direct and feature-based scan matching reveal the necessity for their combination to provide more robust estimation in the information-deprived environments. From the benchmarking re-

sults, it can be noticed that our semi-direct approach outperforms the state-of-the-art methods by a considerable margin in various scenarios, achieving the leading results with 68.3% average relative fitness and 5.71 cm average RMSE distance, respectively. Due to the fact that our semi-direct method implements the scan matching in a coarse-to-fine manner, it is less sensitive to unmodelled artifacts such as moving objects or undergoing the view occlusion, viewpoint changes, and information-deprived environments. For instance, it can be seen from Fig. 3.7 that a van constantly appear in front of the ego-vehicle for the KITTI sequence 04. The existence of moving objects in the scene degrades a lot the performance of both direct and feature-based ICP methods, see the fifth row of Tab 3.2. In this case, the D-ICP and F-ICP tend to be stuck in the local minima, which may partially align the point clouds and provide low fitness registration results. Indeed, the undergoing environments and realized ego-trajectories have a deep impact on

Figure 3.7 – Moving objects in the scene degrade the state estimation performance for the KITTI sequence 04

the performance of LiDAR scan matching. To be more specific, our approach obtains promising results in the highway scenarios (KITTI 01), which is very challenging for direct and feature-based methods because of the relatively low frame rate compared with the high-speed ego-motion(see the second row of Tab 3.2 and Tab 3.3). The adaptive initialization strategy in Eq. 3.8 ensures that the starting point for the pose optimization is not biased. With a reliable initialization point considering the last pose heuristics, it takes fewer iterations for dense point-cloud alignment to converge, and it is also more likely to obtain the global minima even for the fast motion (see Fig. 3.6(b)).

3.2 3D LiDAR Scan Matching Uncertainty Modeling

For LiDAR scan matching, it is crucial to evaluate the corresponding uncertainty for the estimated relative poses. This is beneficial for fusing the estimates with other

sensors or for assigning weights to the nodes in a pose-graph to reduce the local errors. Another advantage of uncertainty modeling is to bound the estimation error within a known confidence interval, which is the prerequisite for roads obstacle avoidance and interactive navigation tasks.

There exit several methods like in [86] and [87] for the scan matching covariance estimation. A closed-form covariance estimation method is presented in [86], which lay on the basis of objective function linearization around the optimal estimation. In order to correctly propagate the uncertainty from the measurement space to the estimation domain, the second-order derivatives of the objective function are calculated and applied to the initial measurement noise. Nevertheless, it needs to be noted that the closed-form covariance estimation only considers the uncertainty caused by the sensor noise; thus, it could not apply to the local minima situations. Monte Carlo simulation [87] is another branch of the covariance estimation, which iteratively contaminates the LiDAR scan and the initial guess with white noise. Then, several pose estimates under different conditions can be computed. Based on Monte Carlo sampled scan matching results, the distribution of the relative pose estimation can be reconstructed. Nevertheless, the brute force sampling is time-consuming and limited in its application scenarios.

In order to maintain the computation efficiency and estimation accuracy, the derivative-free covariance estimation method in [88] is adopted to assess the scan matching quality. Instead of indoor pose confidence estimation with the RGB-D sensor [88], it is reformulated in this thesis to predict the LiDAR scan matching uncertainty in large-scale outdoor scenarios with noise and large viewpoint changes. Given the two consecutive LiDAR scans in the local sensor frame ${}^{\mathbf{T}_t}\mathbf{S}_t$ and ${}^{\mathbf{T}_{t+1}}\mathbf{S}_{t+1}$, the estimated relative pose ${}^{t+1}\hat{\mathbf{T}}_t$ tightly aligns the corresponding points[3] $\{\mathbf{x}_i\} \in \mathbf{S}_t$ and $\{\mathbf{x}_i^{'}\} \in \mathbf{S}_{t+1}$ in the point clouds. Essentially, the uncertainty estimation is based on the inconsistency indicator $\mathbf{D}({}^{t+1}\hat{\mathbf{T}}_t, \mathbf{T}_t, \mathbf{T}_{t+1})$ for all valid pairwise correspondences $(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}$ within a predefined distance threshold $\left\| {}^{t+1}\hat{\mathbf{T}}_t \mathbf{x}_i - \mathbf{x}_i^{'} \right\| < \epsilon$. The error tolerance ϵ is linked to the sensor precision, which is set as 0.1 m in this thesis. And the information matrix can then be extracted via the local parameterization $\xi = (\mathbf{r}, \mathbf{t}) = (\alpha, \beta, \gamma, x, y, z)$[4] about the estimation and true value

3. $\{\mathbf{x}_i\}$ are points with the homogeneous coordinates and $\{\mathbf{x}_{i_{vec}}^{\wedge}\}$ are skew symmetric matrices of points with the vectorized cartesian coordinates

4. For the local parameterization of transformation discrepancy, β is far away from its singular position of $\frac{\pi}{2}$, thus the gimbal lock issue of Euler angle parametrization is avoided

discrepancies $\mathbf{T}_t^{-1}\mathbf{T}_{t+1}{}^{t+1}\hat{\mathbf{T}}_t$ in such manner:

$$\mathbf{T}_t^{-1}\mathbf{T}_{t+1}{}^{t+1}\hat{\mathbf{T}}_t \approx \begin{bmatrix} \mathbf{I}_3 + \mathbf{r}^\wedge & \mathbf{t} \\ 0 & 1 \end{bmatrix} \tag{3.11}$$

After that, the inconsistency indicator can be approximated as:

$$\begin{aligned} \mathbf{D}({}^{t+1}\hat{\mathbf{T}}_t, \mathbf{T}_t, \mathbf{T}_{t+1}) &= \sum_{(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}} \left\| {}^{t+1}\hat{\mathbf{T}}_t \mathbf{x}_i - \mathbf{x}_i^{'} \right\|^2 \\ &= \sum_{(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}} \left\| {}^{t+1}\hat{\mathbf{T}}_t \mathbf{x}_i - \mathbf{T}_{t+1}^{-1}\mathbf{T}_t \mathbf{x}_i \right\|^2 \\ &= \sum_{(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}} \left\| \mathbf{T}_t^{-1}\mathbf{T}_{t+1}{}^{t+1}\hat{\mathbf{T}}_t \mathbf{x}_i - \mathbf{x}_i \right\|^2 \\ &\approx \sum_{(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}} \left\| \mathbf{r} \times \mathbf{x}_{i_{vec}} + \mathbf{t} \right\|^2 \\ &= \sum_{(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}} \left\| \begin{bmatrix} -\mathbf{x}_{i_{vec}}^\wedge & \mathbf{I_3} \end{bmatrix} \xi \right\|^2 \\ &= \xi^T \Lambda \xi \end{aligned} \tag{3.12}$$

where Λ is the information matrix in the quadratic form with the expression as follows:

$$\Lambda = \sum_{(\mathbf{x}_i, \mathbf{x}_i^{'}) \in \mathbf{C}_{ii'}} \begin{bmatrix} -\mathbf{x}_{i_{vec}}^\wedge & \mathbf{I_3} \end{bmatrix}^T \begin{bmatrix} -\mathbf{x}_{i_{vec}}^\wedge & \mathbf{I_3} \end{bmatrix} \tag{3.13}$$

The information matrix gives a direct reflection of the tightness of two LiDAR scans being aligned, and its inverse Λ^{-1} can be considered as the covariance matrix to model the scan matching uncertainty. Compared with the scan alignment itself, the covariance matrix is calculated only once at the final iteration, and its estimation time is negligible since it is derivative-free. The validity of the proposed approach is validated through the publicly available KITTI dataset sequence 05, which contains road intersection and dynamic environments. (See Fig. 3.8 and Fig. 3.9)

The pose uncertainty estimation helps to bound the pose error within a known confidence interval. On the one hand, it can be seen from Fig. 3.8 that our predicted confidence interval could accurately bound the estimation errors during most of the time. It is attributed primarily to the pre-conducted outdoor ground points removal, scattered outliers removal, and multi-level semi-direct scan matching that incorporate more reliable

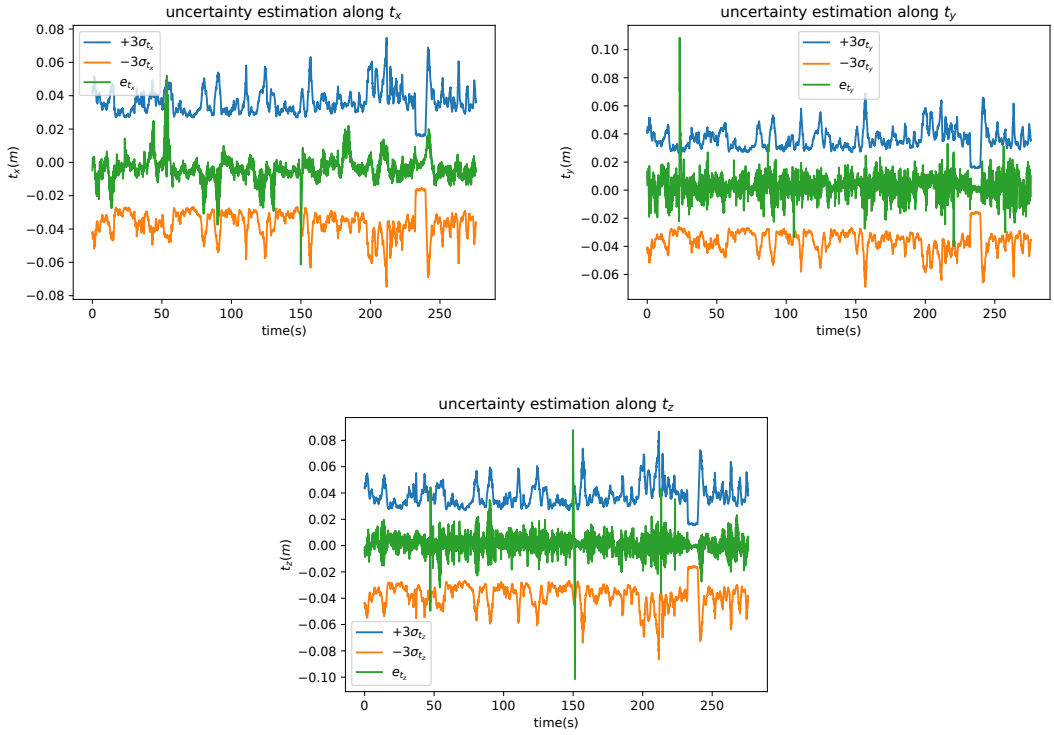


Figure 3.8 – Scan matching uncertainty estimation along t_x, t_y and t_z for the KITTI sequence 05

correspondences for the pose estimation. As it can be inferred from Eq. 3.13, more valid correspondences will result in a more confident state estimation, which is in line with the principle of maximum likelihood estimation. It can also be seen from Eq. 3.13 that the farther the inliers $\{\mathbf{x}_i\}$ locate from the local sensor frame, the more confidence we gain from the pairwise correspondences. On the other hand, it is observed that few overshoots occur at the road intersection with some moving objects in the scene, see Fig. 3.9. The ego-motion at the road intersection will cause inadequate overlapping of subsequent LiDAR scans for state estimation, and the scan matching convergence can not be guaranteed in this situation. Besides, the existence of the moving object further complicates the point cloud registration process and may lead to inconsistency in uncertainty estimation.

To rectify the misaligned point clouds, the pose-graph global optimization technique proposed in [89] is applied to reduce the local drifts, see Fig. 3.10. A pose graph is formulated with nodes representing the poses and edges representing pairwise registration, and then the deviated poses are penalized and smoothed with the global pose-graph opti-

(a) Frame 108 in the KITTI sequence 05

(b) Frame 500 in the KITTI sequence 05

(c) Frame 1514 in the KITTI sequence 05

Figure 3.9 – Partially overlapped point clouds at the road intersection degrade the scan matching performance which corresponds to the overshoots in Fig. 3.8

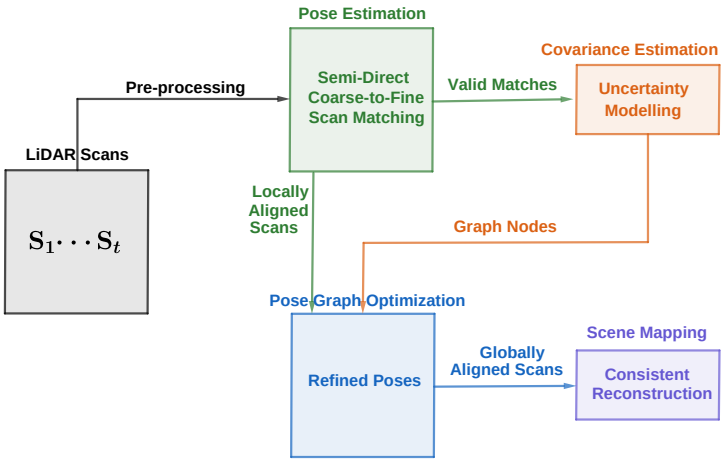
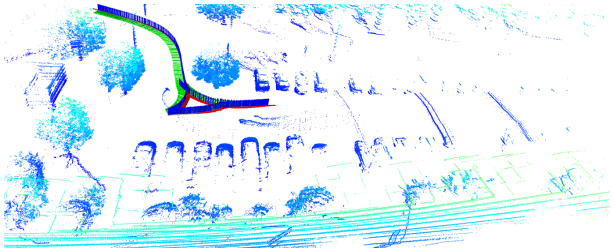


Figure 3.10 – Overview of the proposed uncertainty modeling integrated in the pose graph optimization

mization. The results in Fig. 3.11, indicate that pose-graph optimization can significantly improve the pose estimation accuracy and scene mapping consistency, where the point clouds are well registered with high fidelity.

(a) The main parking scene visualization with the front camera



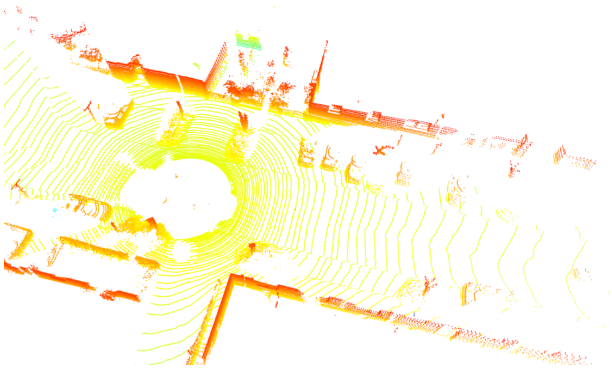
(b) The main parking scene reconstruction with the VLP-16 LiDAR

Figure 3.11 – The main parking area scene mapping and self-positioning with the VLP-16 LiDAR at the Centrale Nantes Campus (LS2N 00)

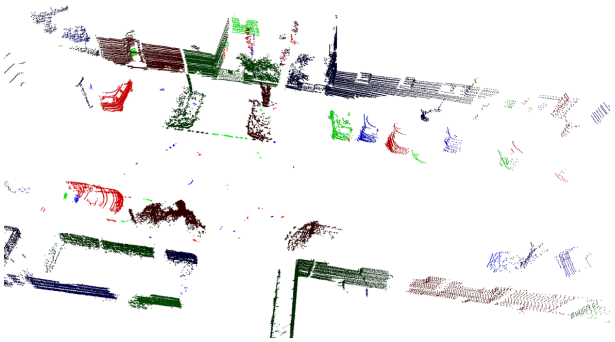
3.3 3D LiDAR-based Objects Detection

For the intelligent vehicle perception system, identifying the surrounding objects plays an important role for the decision-making. Compared with the 2D objects detection on the image plane, 3D objects detection with the LiDAR sensor has the advantages of better object size recognition and spatial range prediction. In this section, both the unsupervised and supervised object detection approaches will be detailed and discussed.

Before the era of deep learning with neural networks, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [90] is a well-known approach for point cloud clustering and objects detection. The DBSCAN clustering algorithm aims to recursively classify the points into the same group when are within the given clustering search radius. The points that do not belong to any cluster are labeled as noise to be filtered out. The two hyper-parameters for the DBSCAN clustering are ϵ and n, which are the radius to aggregate the neighboring points and the minimum number of points required to form the dense clusters. The DBSCAN clustering algorithm is superior to the conventional KMeans method [91] since it can effectively recognize arbitrarily formed clusters without knowing the number of clusters beforehand. In order to efficiently segment the 3D objects



(a) The kitti HDL-64E LiDAR point cloud visualization

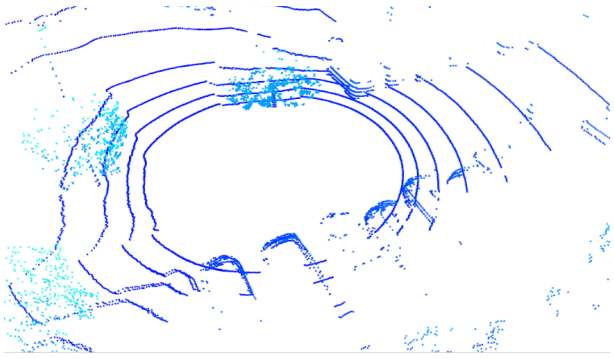


(b) The objects detection results with the DBSCAN clustering algorithm for the kitti dataset

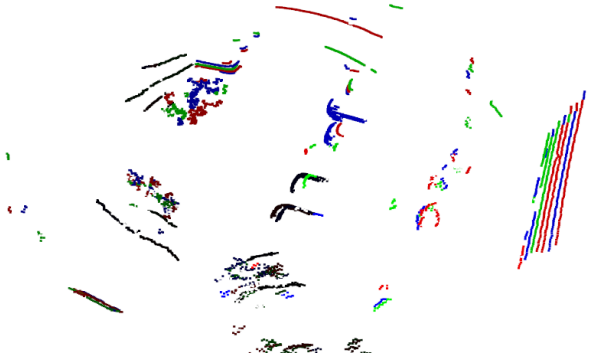
Figure 3.12 – The 3D objects detection from the kitti HDL-64E LiDAR point cloud with the DBSCAN clustering algorithm

from the background scene for the intelligent vehicle application, the point cloud pre-processing such as voxel down-sampling and ground points removal in the Section 3.1 are indispensable. Nonetheless, it needs to be noted that the DBSCAN clustering is essentially an unsupervised algorithm and it can only segment the scene into several regions according to the point cloud density distribution. As we can see from the Fig. 3.12, the parked cars are labeled with different colors for objects clustering. However, the tree brands and vertical walls, which are not considered as traffic participants, are also segmented from the background. The clustered regions lack semantic labels, which makes the scene understanding and decision-making tasks not trivial. Moreover, the detection results degrade a lot when the point cloud is sparse with the VLP-16 LiDAR, see Fig. 3.13. Therefore, the unsupervised objects detection results can only serve as the preliminary baseline and are not suitable for the real applications.

In order to overcome these difficulties, substantial research efforts have been made to



(a) The LS2N VLP-16 LiDAR point cloud visualization



(b) The objects detection results with the DBSCAN clustering algorithm for the LS2N dataset

Figure 3.13 – The 3D objects detection from the LS2N VLP-16 LiDAR point cloud with the DBSCAN clustering algorithm

learn the 3D feature representation and topological connection from LiDAR point clouds for the end-to-end 3D objects detection. The learning-based method is more applicable than the clustering-based method for different scenarios with different traffic participants. After the training process, the learning-based method has no hyper-parameters to tune in the application which is superior to the clustering-based method. Moreover, the object labels can also be identified during neural inference, which facilitates the contextual understanding of the scene. PointNet [92] is a pioneering unified neural network pipeline that transforms the unordered LiDAR point clouds into a canonical representation, where the encoded point-wise features are invariant to certain transformations. Then, the global feature vector can be obtained after the symmetric max pooling function. And the final segmentation and classification heads which are made of the Fully Connected Network (FCN) layers, take in the global feature vector and provide the ultimate detection out-

puts. VoxelNet [93] is another milestone for efficient 3D objects detection, which extracts feature vector representation directly from equally partitioned 3D voxel grids. The novel Voxel Feature Encoding (VFE) layer allows inter-point interaction within a voxel, which helps to characterize the local 3D shape information. After that, the convolutional middle layers aggregate the local voxel features and enrich the descriptive volumetric representation. At last, a Region Proposal Network (RPN) is responsible to generate the 3D objects detection results. However, the inference time of VoxelNet is too long to be deployed in real-time systems. PointPillars [94] is also a representative approach that quantizes the original point cloud into vertical pillars instead of voxels representation for the efficient feature extraction. The pillar-wise encoding of features avoids the computationally expensive 3D convolutions and 2D convolutional backbones can be directly applied to the pseudo image for the high-dimension neural inference. For the detection head, the Single Shot Detector (SSD) [95] is adopted to regress the orientated 3D bounding boxes along with their semantic labels, which enables inference at 62 Hz. The benchmarks results shown in [94] suggest that PointPillars is an appropriate encoding for objects detection in point clouds and the sample detection result can be seen from Fig. 3.14. Neverthe-

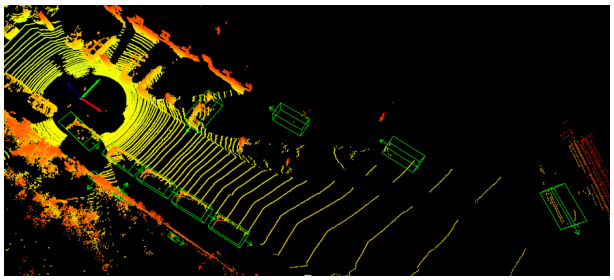


Figure 3.14 – The PointPillars 3D objects detection reuslts with the kitti HDL-64E LiDAR point cloud

less, since the 3D bounding boxes usually have different dimensions and orientations, the anchor-based 3D detectors may encounter the difficulty in regressing the 3D oriented bounding boxes from the axis-aligned 2D ones. We can observe from Fig. 3.14 that the orientation of the predicted bounding boxes are a bit biased due to the predefined bounding boxes anchors. Thus, the CenterPoint framework [96] is proposed to align the bounding boxes with the center-based representation. The CenterPoint network follows the well-known encoder-decoder architecture pipeline, where the point clouds height and intensity information are encoded in the Bird-Eye-View (BEV) map representation. A resnet-based backbone is used to extract features on the flattened BEV images, which

is followed by the center heatmap head and property regression heads. Specifically, the center heatmap head helps to locate the object centers and infer their semantic classes. And the attributes of objects' 3D size and yaw orientation can be obtained from the up-scaled feature maps with the property regression heads. The outputs of the CenterPoint network are expressed as $\{(C_x^i, C_y^i, C_z^i), (L^i, W^i, H^i), \theta^i, S^i\}_{i=1...n}$, which represent center coordinates, dimensions, heading angles and semantic classes (vehicles, cyclists or pedestrians) of n detected objects respectively. The center-based 3D object detection captures real-scale range and shapes of the objects through rotationally invariant points, which is more robust during the ego-turning phases, see Fig. 3.15.

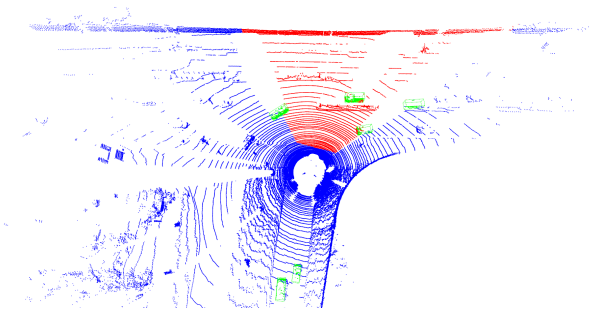


Figure 3.15 – The LiDAR-based 3D object detection results with the CenterPoint network at the road intersection. The red LiDAR points are in the front camera field of view [96]

3.4 Summary

In this chapter, for the sake of overcoming the weakness of individual LiDAR-based direct or feature-based methods, we propose a hybrid semi-direct scan matching approach in Section 3.1 to guarantee convergence in challenging environments such as undergoing the high-speed motion and traveling in repetitive, feature-less environments. On this basis, extensive registration results on city, residential, highway, and parking scenarios with the relative fitness and RMSE metrics are presented and discussed. The superiority of the proposed semi-direct LiDAR-based scan matching method is validated with both HDL-64E and VLP-16 velodyne LiDARs. It is demonstrated that the proposed approach outperforms the state-of-the-art and achieves the leading results with 68.3 % average relative fitness and 5.71 cm average RMSE, respectively. At the same time, scan matching uncertainty is modeled in Section 3.2 as well to evaluate the final convergence accuracy. Furthermore, we also analyze the possible sources that may lead to scan matching diver-

gence in various scenarios. It is demonstrated that ground points and dynamic objects such as vehicles or pedestrians are the main causes of pose estimation accuracy decrease. It is also noticed that significant errors frequently occur near the road intersections and in highway scenarios, where it is more likely to come across the dynamic vehicles and the geometric information is not adequate for reliable state estimation. Thus, we explore the unsupervised and supervised methods for 3D objects detection in Section 3.3. As our current approach only exploits the geometric cues for pose estimation, we will focus on leveraging semantic information to develop more efficient outliers rejection mechanisms, which further ameliorates the scan matching accuracy and robustness with the presence of several moving objects in the following Chapter 4.

Chapter 4

LiDAR-Vision Fusion-based Perception and State Estimation

In the previous Chapters, the visual and range sensors are employed individually for the environment perception and state estimation. In this thesis, environment perception specifically refers to moving objects segmentation and state estimation stands for concurrent ego-pose estimation and map points registration. The Light Detection and Ranging (LiDAR) and the camera sensors are commonly employed to measure the changes in the environment, on which basis, the high-level tasks such as object detection, ego-motion estimation and obstacle avoidance could be performed. The strength and weakness of individual visual and range sensors have been discussed in Section 1.1. And the complementary features of the visual and range sensors encourage us to adaptively combine them for robust perception and state estimation, which efficiently compensates for the individual sensing modality weakness. Sensor fusion allows different sensors to work collaboratively and enables a more reliable perception of the surroundings, which facilitates the full autonomy for the intelligent vehicles.

4.1 Semantic-Guided LiDAR-Vision Moving Objects Segmentation

With the advent of deep learning, object detection based on neural networks could be applied to predict the object position and class on the image plane [42] and 3D point clouds [97] [96] in an end-to-end manner. These neural inference frameworks are mature and can achieve the real-time performance for object detection, which facilitates the on-board integration. Nonetheless, for the objects with the same class, their states of motion (static or dynamic) are not distinguished during the neural inference. Indeed, moving objects are considered as the most unstable traffic participants, which will corrupt the ego-

motion estimation and mapping process. Thus, more attention should be given to them when designing an intelligent vehicle system. In this Section, more attention is paid on the problem of moving objects segmentation, which is primarily concerned with the traffic participants such as pedestrians, cyclists and vehicles. A LiDAR-Vision fusion approach is proposed to combine the LiDAR-based semantic cues and vision-based geometric clues to identify the objects position and state of motion jointly.

The visual sensors provide dense texture and color information of the scene, which facilitates the geometric correspondence establishment and contextual understanding. In [98], a purely geometric mono-vision based approach is proposed for moving objects segmentation in challenging urban areas. The epipolar, trifocal tensor, and structure consistency constraints are flexibly combined to classify the pixel-wise non-static points. The dynamic pixels are then clustered by the connected components labeler for instance-level moving objects segmentation. The optical flow consistency analysis also helps to segment the moving objects from the static background. In [99], the dynamic objects are identified with optical flow-based point trajectories clustering and these moving objects are then excluded from dense SLAM estimation in dynamic environments. In [100], the Flow Vector Bound (FVB) constraint is combined with graph-based clustering for incremental motion segmentation. Nonetheless, the aforementioned methods only leverage geometric information for the object clustering, which often fails in complex scenes. A novel Semantic-Guided RANSAC approach is thus presented in [101] for moving objects segmentation in heavy traffic scenarios. The semantic constraint provides potential moving objects prior and the geometric epipolar residuals are used for the final moving objects verification, which exhibits promising results. Despite of the efficient moving objects segmentation on the visual image plane, the scale metric remains ambiguous, which can be solved by the integration of stereo vision system [102] or range-based sensors [103]. For LiDAR-based perception and ego-motion estimation, robust kernels are commonly adopted to ease the negative impact of outliers. The identified outliers could then be clustered to construct the moving objects in the scene. It is shown in [104] that, the outlier filters such as Tukey, Huber and Cauchy kernels could greatly mitigate the outliers effect for point clouds registration. Besides, the data-driven ResNet50-based method is proposed in [105] to infer the point-wise probability of being dynamic with only a single frame. On this basis, the scene reconstruction module takes the network output of dynamic objects probability for static components mapping. Then, the SpSequenceNet is designed in [106] to operate directly on 4D point clouds (consecutive 3D point clouds) for moving objects segmenta-

tion. Both the spatial and temporal information of LiDAR point clouds are exploited to extract the motion status. However, the SpSequenceNet training and prediction are computationally intensive due to the massive point clouds size. Recently, an innovative range image-based algorithm named Removert is presented in [107]. In the Removert framework, the dynamic objects are pruned from the query LiDAR scans via scan-to-map consistency check. Meanwhile, the pre-built map is corrected with the multi-scale false prediction reverting. As the prior map is not trivially accessible, the map-free method in [108] inputs the inter-frame range-image residuals to the semantic segmentation networks for the real-time class-agnostic moving objects segmentation. Nonetheless, the end-to-end network needs the ground truth binary masks for training that are quite time-consuming to prepare and refine. In order to overcome the individual sensor limitations, the hybrid methods which take advantage of the LiDAR and visual sensors are proposed in [103] [109] [110] [111]. The stereo vision systems are adopted in [103] to improve the object detection and tracking results of the LiDAR-based perception. Specifically, the vision-based system confirms the surrounding objects existence and their dynamic behavior are better modeled due to the dense visual measurements. In [109], the vision-based segmentation result is fused with the planar LiDAR-based prediction, which achieves an improving 2D Intersection-Over-Union (IOU) rate on the Bird-Eye-View (BEV) plane. Besides, the RGB images are converted to a polar-grid representation in [110], which augments the LiDAR point clouds with the color information for the efficient semantic segmentation. A novel architecture to fuse the precise LiDAR depth information and ERFNet-based visual semantics is presented in [111], which is shown to obtain satisfying objects segmentation results both on the image and BEV plane.

In this section, the proposed LiDAR-Vision fusion approach for real-time moving objects segmentation is detailed. Its overall pipeline is shown in Fig. 4.1. To start with, the LiDAR measurements are used to extract the 3D regions of interest (ROI) for movable objects prediction. Then, with the given calibration parameters, 2D ROI on the image plane can be generated via the 3D-2D perspective projection. In order to determine the state of motion for the potentially moving objects, the temporal consistency check is conducted via the optical flow tracking and epipolar geometry. Subsequently, the instance-level moving objects are back-projected in the LiDAR point clouds for 3D moving objects detection.

Both the visual camera and LiDAR sensors can be applied to detect objects in the scene, with only different description formats for the detected targets. Generally, the

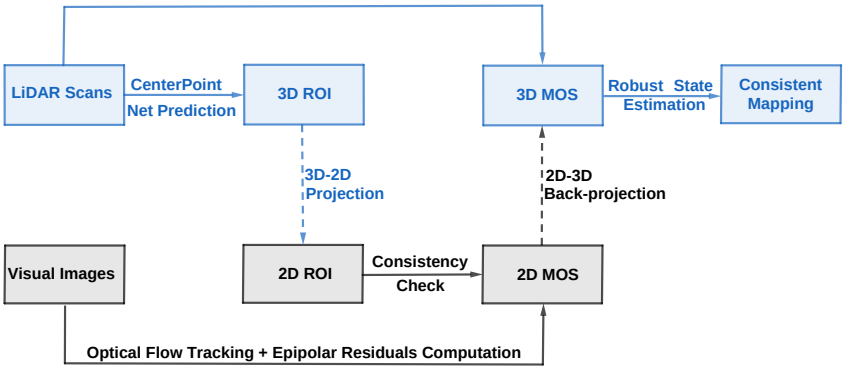


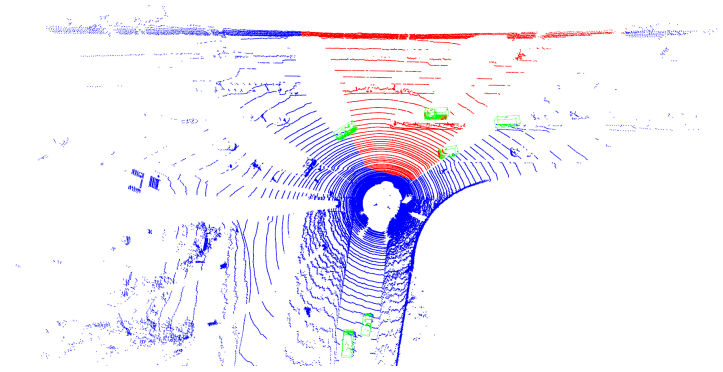
Figure 4.1 – Overview of the proposed LiDAR-Vision fusion approach for moving objects segmentation and state estimation

multi-layer LiDAR sensor is more advantageous for the objects detection task owing to its wider FOV, precise night-vision and long-range perception capabilities. In the proposed approach, the center-based framework CenterPoint [96] is chosen to locate the movable objects and identify their semantic classes in the LiDAR point clouds. The motivation to leverage the CenterPoint framework for movable objects prediction is detailed in Section 3.3. However, due to the inherent LiDAR emission mechanism and resolution issues, the 3D LIDAR point clouds will become increasingly sparse when the objects are far away from the local sensor frame. This makes the state of motion decision-making extremely difficult and drives the adoption of the sensor fusion approach to improve the detection results. In order to develop the LiDAR-Vision fusion approach, it is essential to transform the LiDAR-based detection results from the LiDAR frame $\mathcal{F}_L$ to the image frame $\mathcal{F}_I$ (see Fig. 4.2). It is assumed that the sensors are well synchronized and pre-calibrated with known extrinsic and intrinsic parameters. The perspective projection first takes eight corners ${}^{\mathcal{F}_L}\{\mathbf{x}_i\}_{i=1\ldots8}$ of the 3D bounding box expressed in $\mathcal{F}_L$, and left-multiplies them with the LiDAR-Camera rigid transformation matrix ${}^{C}\mathbf{T}_L$ and image projection matrix ${}^{I}\mathbf{P}_C$ sequentially to get the corner coordinates in $\mathcal{F}_I$.

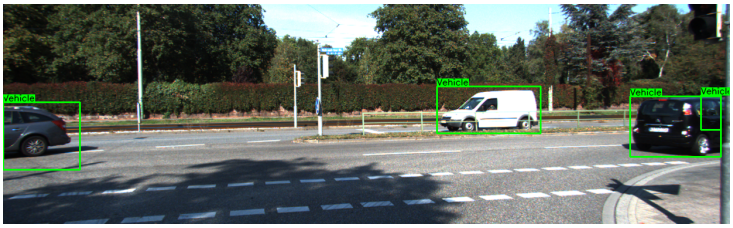
$$ {}^{\mathcal{F}_I}\mathbf{x}_i = {}^{I}\mathbf{P}_C \cdot {}^{C}\mathbf{T}_L \cdot {}^{\mathcal{F}_L}\mathbf{x}_i \tag{4.1} $$

Then the 2D bounding box boundaries can be extracted trivially from the span of the corner coordinates ${}^{\mathcal{F}_I}\{\mathbf{x}_i\}_{i=1\ldots8}$. It is notable that the perspective projection constructs the one-to-one mapping correspondences for the 3D bounding box in $\mathcal{F}_L$ and 2D bounding box in $\mathcal{F}_I$, which provides the possibility for the bounding box back-projection. And

during the 3D-2D perspective projection, the semantic labels of detected objects remain unchanged.



(a) CenterPoint-based 3D objects detection result



(b) The projection of the detection results on the image plane

Figure 4.2 – The LiDAR-based 3D objects prediction and the corresponding projection on the image plane

Given the predicted regions of interest with the LiDAR measurements, the visual multi-view geometry provides a sanity check for the moving objects segmentation validation. To start with, the Shi-Tomasi corners features [28], which remain invariant under the rotation, translation and scaling operation, are detected in the image frame. Then, the detected features are associated with the pyramidal Lucas-Kanade optical flow tracking [54] between two consecutive image frames. And the optical flow backward check is also implemented to reduce the risks of mismatching. During the optical flow-based tracking, if the features lying on the movable objects (rendered in green, shown in Fig. 4.2) are not semantically consistent across two frames, they will be directly identified as dynamic points. After that, the matched features which belong to the background (rendered in blue, see Fig. 4.3), are used to estimate the fundamental matrix $\hat{\mathbf{F}}$ within the RANSAC framework [6]. Since the movable objects points (rendered in green, see Fig. 4.3) are excluded from the estimation, the RANSAC process will converge quickly and provide a reliable fundamental matrix estimation. With the paired background points $(\mathbf{x}_i, \mathbf{x}'_i)_{i=1...n}$

and the estimated fundamental matrix $\hat{\mathbf{F}}$, the corresponding epipolar line $\mathbf{l}'_\mathbf{i} \sim \hat{\mathbf{F}}\,\mathbf{x}_i$ with reduced coefficients $[a_i, b_i, c_i]^T$ can be obtained. The Signed Epipolar Distance (SED) d_i^{SED} from the point $\mathbf{x}'_i = (u'_i, v'_i)$ to line $\mathbf{l}'_\mathbf{i}$ is calculated as:

$$d_i^{SED} = \frac{a_i u'_i + b_i v'_i + c_i}{\sqrt{a_i^2 + b_i^2}} \tag{4.2}$$

Assuming that the measurement noise is normally distributed, then the calculated signed epipolar distance $\{d_i^{SED}\}_{i=1\ldots n}$ will follow the Gaussian distribution as shown in Fig. 4.4, which lays the basis for outlier rejection. As the fundamental matrix transformation compensates the inter-frame ego-motion on the image plane, the static points will have close to zero (noise corruption) SEDs. On the contrary, the points on moving objects tend to get the SEDs exceeding the sigma rule bounds, which will be classified as outliers and segmented from the static background. Nonetheless, it also needs to be mentioned

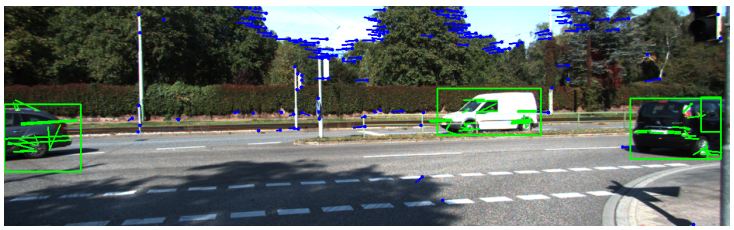


Figure 4.3 – The background corner points (rendered in blue) are tracked with sparse optical flow and are used for robust fundamental matrix estimation

that for objects following the degenerate motions within the epipolar plane, the epipolar constraint alone is not sufficient. This kind of degenerate motion usually happens when the ego-vehicle is following the moving object forward and constantly maintains the straight-line motion. In this case, the Flow Vector Bound (FVB) constraint [112] can be leveraged to detect such moving points with low SEDs. Given the sequential images, the pixels parallax d_i^{FVB} of paired points $(\mathbf{x}_i, \mathbf{x}'_i)$ between two consecutive frames can be computed as:

$$\begin{aligned} \mathbf{x}'_i - \mathbf{K}\mathbf{R}\mathbf{K}^{-1}\mathbf{x}_i &= \frac{1}{z}\mathbf{K}\mathbf{t} \\ d_i^{FVB} &= \left|\mathbf{x}'_i - \mathbf{K}\mathbf{R}\mathbf{K}^{-1}\mathbf{x}_i\right| \end{aligned} \tag{4.3}$$

where the scalar z represents depth value and the matrices $\mathbf{K}$, $\mathbf{R}$ and $\mathbf{t}$ stand for the camera intrinsics, inter-frame rotation and translation respectively. With a set of matched

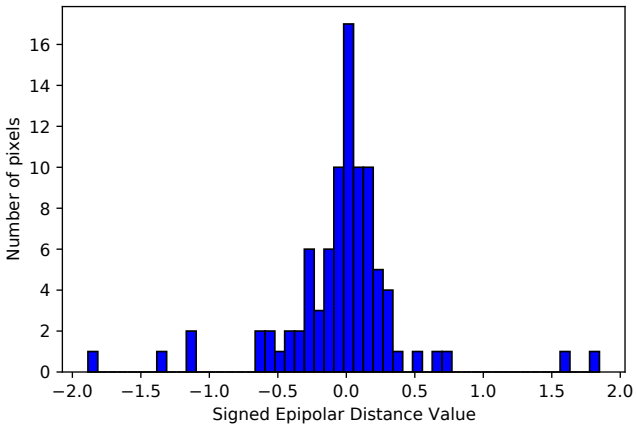


Figure 4.4 – The Signed Epipolar Distance distribution of background points (rendered in blue)

background points $\{(\mathbf{x}_i, \mathbf{x}'_i)_{i=1...n}\}$, their parallax bound $[d_{min}^{FVB},\ d_{max}^{FVB}]$ can be easily found by Eq. 4.3. For the points with degenerate motions, if their parallax value is not within the interval of $[d_{min}^{FVB},\ d_{max}^{FVB}]$, they will also be labeled as dynamic outliers.

The vision-based MOS validation further exploits the underlying geometric and semantic cues, to identify the truly dynamic points lying on the movable objects. The combination of semantic, epipolar and FVB constraints allows for better MOS recognition even in degenerated cases. From the statistical point of view, the likelihood of an object being dynamic depends on proportion of outliers lying inside, on which basis, the instance-level MOS probability can be obtained. In this thesis, the threshold of 50% is set for the instance-level MOS decision making. It means that for each movable object, if there are more than 50% of the points inside are classified as mobile, the object itself will be considered as a dynamic object. Then, the object bounding box 2D-3D back-projection as depicted in Fig. 4.6, is implemented to get the depth information. And all the points inside the truly moving objects (rendered in red, shown in Fig. 4.5) will be cleared for the following robust ego-motion estimation and consistent scene mapping. In brief, 2D-3D instance-level MOS algorithm can be summarized as:

Algorithm 3 2D-3D Instance-Level MOS Algorithm

Input: Predicted Movable Objects ROI $\{\mathbf{R}_i\}_{i=1...r}^{2D}$

Output: Validated MOS $\{\mathbf{S}_i\}_{i=1...s}^{2D,\ 3D}$

1: Detect the Shi-Tomasi corners and track them with LK-optical flow between two consecutive image frames.
2: Estimate the fundamental matrix $\hat{\mathbf{F}}$ with the feature matches belonging to the background.
3: Compute the SED residual distribution $(\mu^{SED}, \sigma^{SED})$ using Eq. 4.2 and the flow vector bound $[d_{min}^{FVB},\ d_{max}^{FVB}]$ using Eq. 4.3 for the background points.
4: Check the motion status of points inside $\{\mathbf{R}_i\}_{i=1...r}^{2D}$ based on the semantic, epipolar and FVB constraints.
5: Classify the movable object $\{\mathbf{R}_i\}^{2D}$ as validated dynamic object $\{\mathbf{S}_i\}^{2D}$ if the proportion of outliers in $\{\mathbf{R}_i\}^{2D}$ exceeds the threshold 50%.
6: Implement the $\{\mathbf{S}_i\}^{2D}$ back-projection to obtain $\{\mathbf{S}_i\}^{3D}$.
7: Exclude the points inside $\{\mathbf{S}_i\}^{3D}$ for the following robust ego-motion estimation and consistent scene mapping.

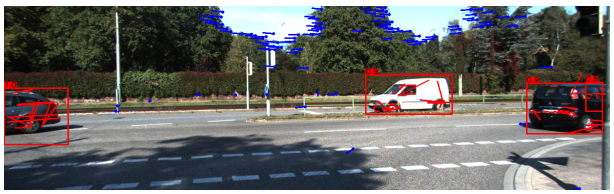


Figure 4.5 – The 2D instance-level moving objects segmentation (rendered in red), along with their probability of being dynamic

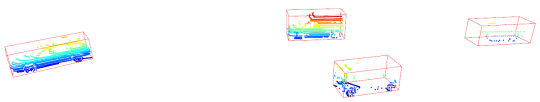


Figure 4.6 – The 3D instance-level moving objects segmentation via back-projection, and all the points inside will be classified as outliers

The ability to segment dynamic components in surrounding environments is essential for the intelligent transportation system. In the proposed approach, the movable objects ROI are predicted with the CenterPoint neural network. Then, the visual multi-view

geometry constraints provide a sanity check for instance-level MOS validation. Such a combination allows for better recognition, which is capable of detecting tiny objects (see Fig. 4.7) and even partially occluded objects (see Fig. 4.5). The right-side vehicle in Fig. 4.5 is occluded on the image plane, which is quite difficult to detect with only visual hints. Nonetheless, the high-resolution LiDAR sensor receives the reflection from part of occluded vehicle and manages to predict its existence, as depicted in Fig. 4.6. Moreover, it is demonstrated in Fig. 4.7 that, the epipolar constraint compensates the vehicle ego-motion which accurately classifies the parked car as static. And the flow vector bound constraint efficiently helps to identify the dynamic vehicle performing degenerate motions on the lane, which facilitates better contextual understanding of the drivable area.

Figure 4.7 – The tiny moving object with degenerate motions (move along the epipolar plane) is successfully segmented with the FVB constraint

Figure 4.8 – The static car parked on the roadside (in blue) and dynamic car driving on the lane (in red) are distinguished and back-projected to the 3D LiDAR scan

4.2 Semantic-Guided LiDAR-Vision Ego-motion Estimation and Scene Mapping

The robust ego-motion estimation is driven by the reliable correspondence matches across frames, where the estimated transformation ${}^{j}\mathbf{T}_{i}$ tends to minimize the overall distance between the paired correspondences $\{(\mathbf{P}_i \Leftrightarrow \mathbf{P}_j) \in \mathbf{M}\}$. In the proposed pipeline, the 3D LiDAR scans are iteratively matched with the point-to-plane metric to deduce the vehicle ego-motion, which reaches centimeter-level precision as shown in Chapter 3. However, the existence of dynamic objects in the scene tends to cause the scan misalignment,

thus degrading the registration accuracy of sequential LiDAR scans. In order to mitigate the impact of outliers in the objective function $\mathbf{F}({}^{j}\mathbf{T}_i)$ minimization, the robust kernel [104] functions $\rho(\mathbf{r}_{ij})$ adaptively adjust the weights of the matches with large residuals, where $\mathbf{r}_{ij}$ is the point-to-line ICP residuals.

$$\mathbf{F}({}^{j}\mathbf{T}_i) = \sum_{(\mathbf{P}_i,\mathbf{P}_j)\in\mathbf{M}} \rho((\mathbf{P}_j - {}^{j}\mathbf{T}_i\mathbf{P}_i)\mathbf{n}_j) \tag{4.4}$$

where $(\mathbf{P}_i \Leftrightarrow \mathbf{P}_j)$ are the correspondences belonging to the matched points set $\mathbf{M}$, and $\mathbf{n}_j$ is the normal vector around $\mathbf{P}_j$ for calculating the point-to-plane distance. Nevertheless, the kernel-based Iterative Closest Point (ICP) method is not sufficient to handle the constant dynamic objects corruption. To solve this problem, the proposed approach distinguishes the instance-level moving objects as in Section 4.1, with the semantic-guided LiDAR-Vision information fusion. With the segmented moving objects back-projected to the 3D LiDAR scans, the weighting coefficients $\rho(\cdot)$ for the paired points lying inside the dynamic objects are uniformly assigned as zero, which further reduces the influence of dynamic objects in challenging scenarios. Moving objects are considered as the most unstable traffic participants, which will corrupt the ego-motion estimation and mapping process. Since the moving objects are not temporally consistent, they do not belong to the permanent components of the scene. Therefore, moving objects should be eliminated from the mapping process in order to build a consistent representation of the scene, as shown in Fig. 4.9. The reconstructed static map will promote high-level tasks such as map-based localization and path planning. In order to align the LiDAR scans in the global frame, the poses from the ego-motion estimation thread will then be leveraged for static point clouds registration.

The effectiveness of the proposed sensor fusion-based ego-motion estimation system is validated with the the KITTI dataset [55]. The experimental evaluations are conducted with the challenging city category sequences [1], which were recorded in heavy traffic hours. The LiDAR-Camera sensor setup is adopted with the known calibration parameters, where the 64-layer Velodyne HDL-64E LiDAR gives accurate range information and the RGB-camera provides more contextual knowledge of the scene. In the KITTI dataset, only the front-view images are provided. So the focus is paid only on the moving objects segmentation and static scene mapping within the visual sensor field of view. The semantic-guided LiDAR-Vision fusion approach efficiently reduces the outliers

1. http://www.cvlibs.net/datasets/kitti/raw_data.php?type=city

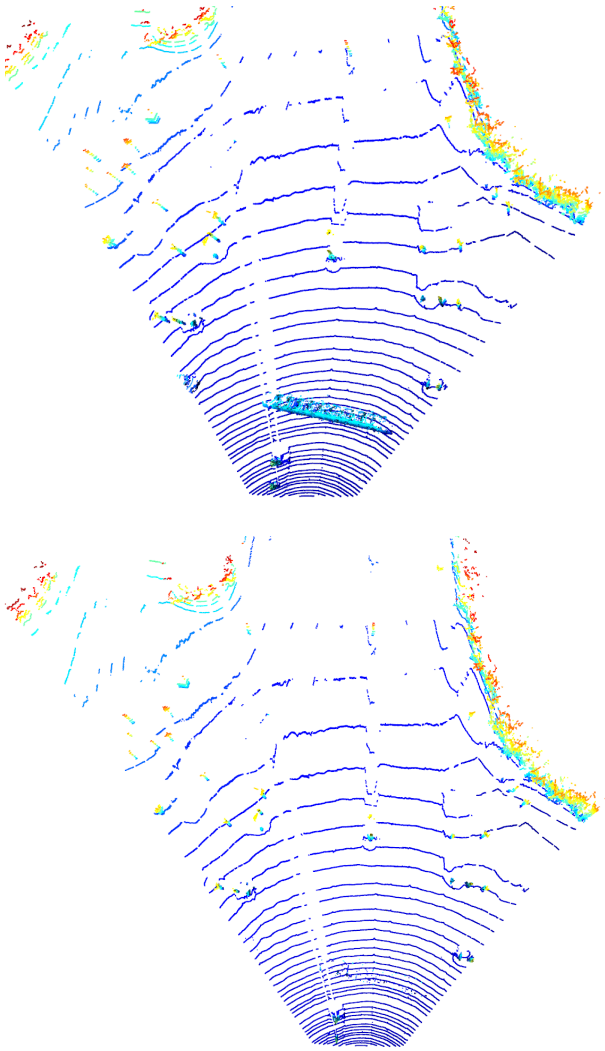


Figure 4.9 – The ghosting effect of a moving car is greatly reduced due to the semantic-guided moving objects segmentation and removal

effect in the 3D LiDAR scan matching. In order to quantify the performance of LiDAR scan matching, the evaluation metrics of Relative Fitness (RF) and Relative Root Mean Square Error (RMSE) of the inlier correspondences $\{\mathbf{I}\}$ are used. They are defined as in the Section 3.1. It is shown in Tab 4.1 and Tab 4.2 that, the proposed semantic-guided MOS approach achieves the leading results of 77.9% average fitness and 7.65 cm RMSE respectively. Since the optimization-based ego-motion estimation is usually built upon the static environment assumption. The presence of dynamic objects in the scene may degrade the ego-motion estimation and complicate the map maintenance task. A two-stage prediction-then-validation pipeline is thus designed to segment instance-level

objects in the scene. It is more efficient than the traditional kernel-based methods, since the analysis concentrates on the ROIs instead of the whole points clouds. Compared to the end-to-end DL-methods, the proposed approach relieves from motion segmentation ground truth annotation and training. The multiple constraints combination also ensures the robustness of dynamic outliers rejection in complex situations, such as handling objects with degenerated motions.

Table 4.1 – Semantic-guided ICP Registration Results Benchmarking with RF(%)

Methods / Dataset	Tukey	Huber	Cauchy	Ours
2011_09_26_13	57.2%	55.6%	55.6%	**57.7**%
2011_09_26_17	94.7%	94.7%	94.7%	**95.7**%
2011_09_26_18	78.9%	79.5%	79.1%	**80.2**%
Average	76.9%	76.6%	76.5%	**77.9**%

Table 4.2 – Semantic-guided ICP Registration Results Benchmarking with RMSE (CM)

Methods / Dataset	Tukey	Huber	Cauchy	Ours
2011_09_26_13	10.18	10.37	10.39	**9.98**
2011_09_26_17	5.73	5.75	5.75	**5.66**
2011_09_26_18	7.37	7.41	7.41	**7.31**
Average	7.76	7.84	7.85	**7.65**

4.3 Loosely Coupled LiDAR-Vision Odometry

Accurate ego-motion estimation and a good knowledge of the surrounding environment are crucial for autonomous driving. On autonomous vehicles, the range-based LiDAR and/or vision-based stereo cameras are commonly deployed to perform the task of ego-motion estimation. The most frequently used sensors (Camera and LiDAR) have their own merits and weaknesses under different working conditions as described in the beginning of Chapter 4. Thus, the fusion of range and visual sensors allows to compensate respective shortcomings and achieves reliable ego-motion estimation results. The main challenge for long-term ego-motion estimation is error accumulation, especially in environmentally degenerate scenarios. The fusion of range and visual sensors could restrict the local uncertainties and allow to confine the odometry drift. ORB-SLAM [9] is

regraded as a typical representative of vision-based SLAM. Oriented FAST and Rotated BRIEF (ORB) features are extracted and matched for real-time ego-motion estimation, and meanwhile the Bag-Of-Words (BOW) [113] dictionary is queried for loop closure and drift cancellation. This method can accurately localize the mobile platform and create a sparse feature map of its surroundings with limited computation resources. ORB-SLAM2 is proposed in [12] with stereo observations based back-end which solves the scale ambiguities for trajectory estimation. LiDAR scan-matching is a fundamental process to estimate platform motion and to create a 3D map with the laser scanner. A popular approach for LiDAR based localization is LOAM [15]. It conducts Iterative Closest Point (ICP) scan-matching for 3D point clouds registration, which is followed by a global scan-to-map alignment in order to reduce local errors. Feature alignment problem can be solved using the well known Levenberg–Marquardt optimizer. Despite the success and popularity of ORB2 SLAM [12] and LOAM [15], they are in fact deterministic algorithms. They do not effectively handle the sources of uncertainty. As a result, they provide overconfident ego-motion estimation results across frames. In this section, a loosely coupled sensor fusion approach is proposed for vehicle localization with range and visual sensors. Measurement uncertainties for visual and range sensors are properly defined for ego-motion estimation. Backward covariance propagation [33] is utilized to transform the covariance from measurement domain to estimation domain. At the same time, forward covariance propagation is leveraged to transform the uncertainty from manifold space to Euclidean space. The covariance intersection filtering [114] enables adaptive fusion of the two sensors given their respective uncertainties.

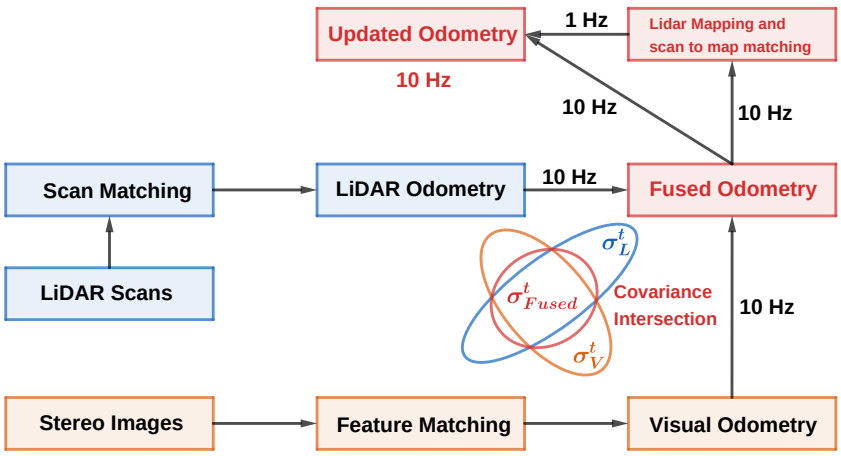


Figure 4.10 – Overview of the proposed loosely-coupled sensor fusion scheme

In Fig.4.10, it can be seen that the proposed sensor fusion framework starts with a descriptor-based visual feature tracking module to estimate vehicle ego-motion. Meanwhile, LiDAR distance-based scan-to-scan matching runs in parallel for ego-motion estimation. Backward covariance propagation transforms the uncertainty from measurement space to estimation space, which helps to obtain the uncertainty of frame-to-frame ego-motion estimation for both sensor modalities. Covariance intersection filtering ensures that the uncertainty of state does not expand after the sensor fusion, which combines two frame-to-frame poses elegantly. Then, the robustified pose is used for LiDAR point cloud registration. Scan-to-map matching afterwards further reduces local drift caused by frame-to-frame estimation. The updated odometry is the final output which is published at 10Hz frequency.

To denote coordinate systems, the convention in the Section 4.1 is followed to use uppercase letter to indicate different coordinate frames. In visualization and sensor fusion steps, the vehicle pose is expressed with 3D translation and RPY Euler angle rotation. However, in order to avoid singularity problem, optimization is made on their manifold with the Lie algebra. In the following, coordinate systems being used are explained.

— Camera sensor coordinate system $\mathcal{F}_{C_t}$ at timestamp t is defined at the camera optical center. The x-axis, y-axis and z-axis point rightward, downward and forward respectively as the camera configuration in [55].

— LiDAR sensor coordinate system $\mathcal{F}_{L_t}$ at timestamp t is defined at the LiDAR scanner center. The x-axis, y-axis and z-axis point forward, leftward, and upward respectively as the LiDAR configuration in [55].

— World coordinate system $\mathcal{F}_W$ is defined as $\mathcal{F}_{C_0}$ which is the initial frame of the camera coordinate system, and lidar-camera extrinsics ${}^{C}\mathbf{T}_{L}$ is assumed to be known beforehand.

In feature-based stereo vision odometry, key points with local descriptors are matched to deduce the camera motion with scale metrics. Providing camera intrinsics $\mathbf{K}$, stereo feature points belonging to the previous frame are triangulated in the first step. And then transformed triangulated points are re-projected via the perspective projection operation $\mathrm{Pr}^{l}(\cdot)$, $\mathrm{Pr}^{r}(\cdot)$ onto the left and right images respectively considering the 6 dof ego-motion estimation variable ${}^{\mathcal{F}_{C_{t-1}}}\hat{\Theta}_{\mathcal{F}_{C_t}}$:

$$
{}^{\mathcal{F}_{I_t}}\hat{\mathbf{x}}_i = \begin{bmatrix} \mathrm{Pr}^{l}\left(\mathbf{K}, {}^{\mathcal{F}_{C_{t-1}}}\hat{\Theta}_{\mathcal{F}_{C_t}}, {}^{\mathcal{F}_{I_{t-1}}}\mathbf{x}_i\right) \\ \mathrm{Pr}^{r}\left(\mathbf{K}, {}^{\mathcal{F}_{C_{t-1}}}\hat{\Theta}_{\mathcal{F}_{C_t}}, {}^{\mathcal{F}_{I_{t-1}}}\mathbf{x}_i\right) \end{bmatrix} \tag{4.5}
$$

where ${}^{\mathcal{F}_{I_t}}\hat{\mathbf{x}}_i = (\hat{u}_{i,l}, \hat{v}_{i,l}, \hat{u}_{i,r}, \hat{v}_{i,r})^T$ is the prediction in the current frame and ${}^{\mathcal{F}_{I_{t-1}}}\mathbf{x}_i = (u_{i,l}, v_{i,l}, u_{i,r}, v_{i,r})^T$ is its correspondence in the previous frame. In general, the optimal relative camera transformation can be estimated by minimizing the weighted squared error of measurements and predictions.

$$ {}^{\mathcal{F}_{C_{t-1}}}\Theta^*_{\mathcal{F}_{C_t}} = \underset{{}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}}{\operatorname{argmin}} \mathbf{F}(\mathbf{x}, {}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}) = \underset{{}^{\mathcal{F}_{C_{t-1}}}\Theta_{\mathcal{F}_{C_t}}}{\operatorname{argmin}} \sum_{i=1}^{N_c} \left\| {}^{\mathcal{F}_{I_t}}\mathbf{x}_i - {}^{\mathcal{F}_{I_t}}\hat{\mathbf{x}}_i \right\|^2_{\Sigma} \tag{4.6} $$

where $\| \cdot \|^2_{\Sigma}$ is the Mahalanobis distance with $\Sigma^{-1}_{\mathbf{x}_i}$ as the information matrix for the i_{th} measurement. To handle estimation parameters that do not belong to Euclidean spaces, the common strategy is to transfer the error minimization to its corresponding manifold. In this thesis, iterative optimization update for estimated parameters is made using Lie algebraic perturbation model [36]. Operator $\boxplus$ is a generalization of the normal addition operator, which is defined as $\delta\epsilon \boxplus \hat{\Theta} \triangleq exp(\delta\epsilon)\hat{\Theta}$, then $\tilde{\mathbf{J}}_i(\hat{\Theta})$ can be written as

$$ \tilde{\mathbf{J}}_i(\hat{\Theta}) = \left. \frac{\partial \mathbf{e}_i(\delta\epsilon \boxplus \hat{\Theta})}{\partial \delta\epsilon} \right|_{\delta\epsilon \to 0} \tag{4.7} $$

As a result, the evenberg-Marquardt algorithm can be applied without considering additional constraint such as rotation matrix orthogonality.

The same way in LiDAR odometry, edge and planar LiDAR points are tracked to recover the LiDAR pose. For each LiDAR scan point, local curvature c is computed to evaluate its smoothness considering the surrounding area. Let $\mathbb{S}$ be a group of points in the vicinity of $\mathbf{x}_i$ in the same scan layer.

$$ c = \frac{1}{|\mathbb{S}| \cdot \left\| {}^{\mathcal{F}_{L_t}}\mathbf{x}_i \right\|} \left\| \sum_{j \in \mathbb{S}, j \neq i} \left({}^{\mathcal{F}_{L_t}}\mathbf{x}_i - {}^{\mathcal{F}_{L_t}}\mathbf{x}_j \right) \right\| \tag{4.8} $$

Edge and planar points are defined based on c values. The edge line constructed by two edge points at previous frame $({}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j, {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_l) \in {}^{\mathcal{F}_{L_{t-1}}}\mathbb{E}$ forms the correspondence of ${}^{\mathcal{F}_{L_t}}\mathbf{x}_i$. ${}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j$ and ${}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j$ are selected according to nearest neighbor criteria and they belong to different scan layers to increase the point-to-line fitting robustness. The planar patch represented by three points at previous frame $({}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j, {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_l, {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_m) \in \mathbb{H}_{t-1}$ forms the correspondence of ${}^{\mathcal{F}_{L_t}}\mathbf{x}_i$. It is assumed that the closest neighbor of ${}^{\mathcal{F}_{L_t}}\mathbf{x}_i$ is denoted as ${}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j$. ${}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_l$, ${}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_m$ are second and third nearest neighbors of ${}^{\mathcal{F}_{L_t}}\mathbf{x}_i$, one

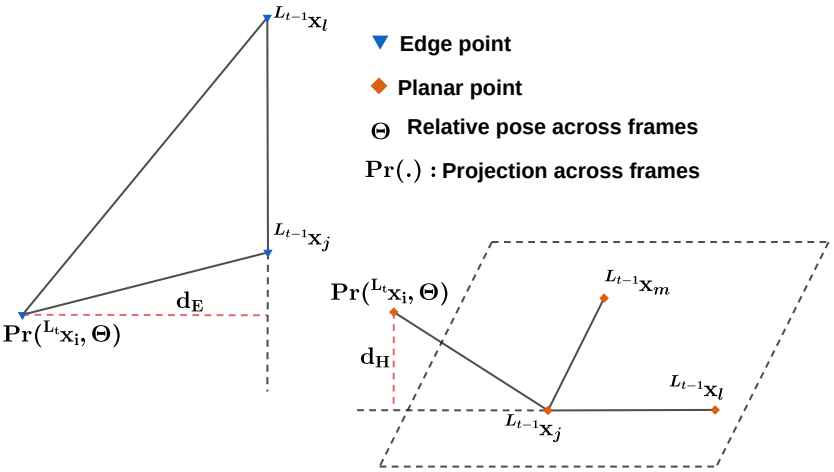


Figure 4.11 – Scheme of edge and planar LiDAR points correspondence projection

belonging to the same scan layer of ${}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j$, and the other in the consecutive scan layer of ${}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j$. With the corresponding relationship of the feature points in hand, according to Fig. 4.11, the distance from a feature point to its correspondence can be calculated, where $|\cdot|$ stands for the norm value.

$$d_{\mathbf{E}} = \frac{\left|\left({}^{\mathcal{F}_{L_t}}\mathbf{x}_i - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j\right) \times \left({}^{\mathcal{F}_{L_t}}\mathbf{x}_i - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_l\right)\right|}{\left|{}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_l\right|} \tag{4.9}$$

$$d_{\mathbf{H}} = \frac{\left|\begin{array}{c}\left({}^{\mathcal{F}_{L_t}}\mathbf{x}_i - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j\right) \\ \left(\left({}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_l\right) \times \left({}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_m\right)\right)\end{array}\right|}{\left|\left({}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_l\right) \times \left({}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_j - {}^{\mathcal{F}_{L_{t-1}}}\mathbf{x}_m\right)\right|} \tag{4.10}$$

The optimal LiDAR relative pose can be obtained by minimizing the weighted sum squared distances of edge and planar points to their correspondences. $\Sigma_{\mathbf{x}_{Ei}^{-1}}$ and $\Sigma_{\mathbf{x}_{Hi}^{-1}}$ stand for the information matrix of $E_{i_{th}}$ edge and $H_{i_{th}}$ planar measurement points and the same optimization strategy is taken as in feature-based stereo vision odometry

$${}^{\mathcal{F}_{L_{t-1}}}\Theta^*_{\mathcal{F}_{L_t}} = \underset{{}^{\mathcal{F}_{L_{t-1}}}\Theta_{\mathcal{F}_{L_t}}}{\operatorname{argmin}} \sum_{E_i=1}^{N_E} d_{E_i}\Sigma_{\mathbf{x}_{Ei}^{-1}} d_{E_i} + \sum_{H_i=1}^{N_H} d_{H_i}\Sigma_{\mathbf{x}_{Hi}^{-1}} d_{H_i} \tag{4.11}$$

Robust ego-motion estimation should be able to provide the uncertainty information associated with the vehicle pose estimates. The sensor fusion phase is driven by the uncertainties in the estimation domain. Thus, the uncertainties coming from visual and

range sensors are analyzed via forward and backward covariance propagation. Although the optimal relative pose can be obtained by minimizing Eq. 4.6, its accuracy also depends on the precision of the corresponding feature points, more specifically, the level of the image pyramid they belong to. The image pyramid [115] is a series of image collections whose resolution gradually decreases in the shape of a pyramid. The image pyramid can be sequentially matched to ensure scale invariance during feature tracking. In this thesis, the image pyramid has 8 levels with the same scale factor 1.2 between two consecutive levels. It is assumed that all points considered in the optimization procedure are well-matched pixel features with only zero mean Gaussian noise $N(0, \Sigma^V_{\mathbf{x}_i})$, with $\sigma_{\mathbf{x}_{u_{i,l(r)}}} = \sigma_{\mathbf{x}_{v_{i,l(r)}}} = 1.2^{level-1}$ as standard deviation for i_{th} measurement. Jacobian matrix $\tilde{\mathbf{J}}_i(\Theta^*)$ is defined in Eq. 4.7, and it converts the uncertainty from measurement space to estimation space. Since the optimize is implemented on manifold, let $\epsilon = log(\Theta^*)$ a 6D vector in the Lie algebra space. Then the Jacobian matrix $\mathbf{J}_{m2e} = \frac{\partial e^\epsilon}{\partial \epsilon}$ is mandatory to propagate the covariance from manifold space to Euclidean space for data visualization and fusion. The uncertainty of frame-to-frame relative pose $\Sigma^V_{\Theta^*}$ can be obtained through Eq. 4.12 and the result is shown in Fig. 4.12.

$$\Sigma^V_{\Theta^*} = \mathbf{J}^V_{m2e}(\sum_{i=1}^{N_c}(\tilde{\mathbf{J}}_i^{V'}(\Theta^*)\Sigma^{V\ -1}_{\mathbf{x}_i}\tilde{\mathbf{J}}_i^V(\Theta^*)))^{-1}\mathbf{J}^{V'}_{m2e} \tag{4.12}$$

In this thesis, a Velodyne HDL-64E is used which provides a ($0° \sim 360°$) azimuth field of view (θ) and ($-24.9° \sim 2°$) elevation field of view (ϕ). According to official velodyne data sheet, range accuracy can reach up to 2 cm which is quite small compared with its range limit 120 m. Hence, each measurement is treated equally and measurement uncertainty $\Sigma^L_{\mathbf{x}_i}$ can be set as identity matrix for each point. Based on such assumption, the uncertainty of scan-to-scan relative pose $\Sigma^L_{\Theta^*}$ can be obtained through Eq. 4.13 and the result is shown in Fig. 4.13.

$$\Sigma^L_{\Theta^*} = \mathbf{J}^L_{m2e}(\sum_{i=1}^{N_E+N_H}(\tilde{\mathbf{J}}_i^{L'}(\Theta^*)\Sigma^{L\ -1}_{\mathbf{x}_i}\tilde{\mathbf{J}}_i^L(\Theta^*)))^{-1}\mathbf{J}^{L'}_{m2e} \tag{4.13}$$

Covariance intersection [114] is a variant of Gaussian process sensor fusion which can combine two estimates under unknown correlations. The covariance intersection combi-

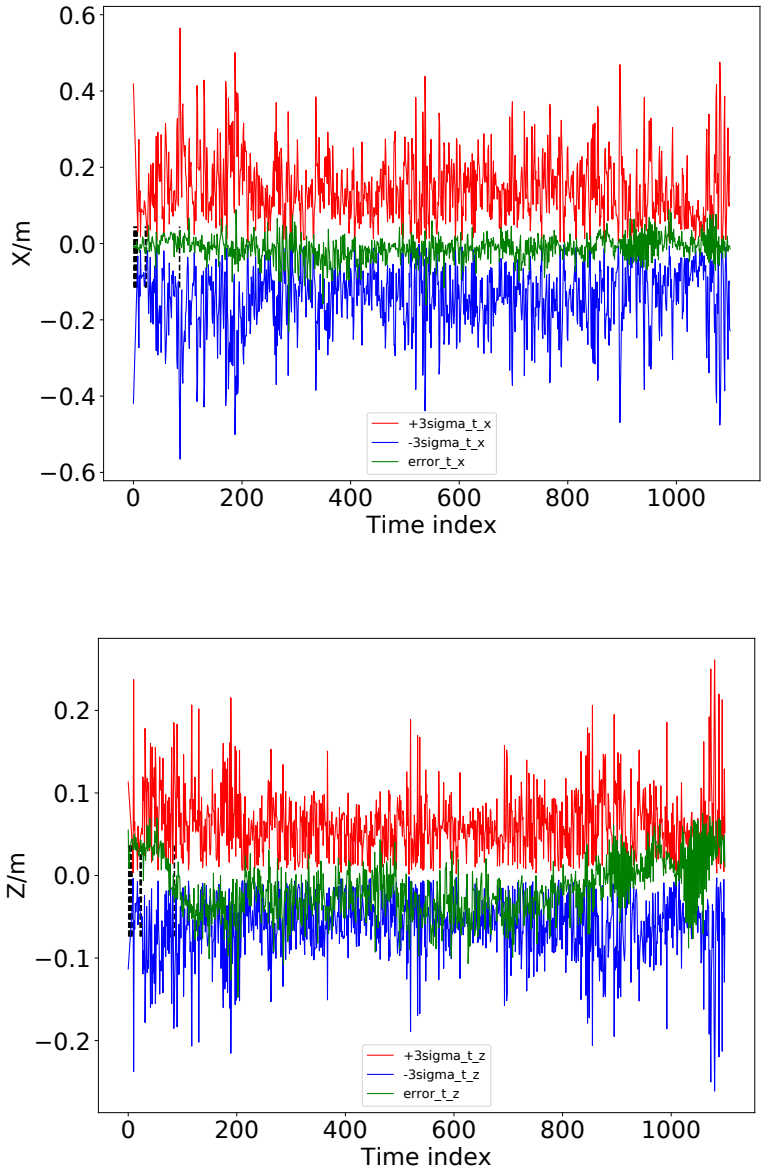


Figure 4.12 – Visual sensor pose estimation uncertainty along t_x and t_z for KITTI sequence 01

nation formulas are given by

$$\begin{aligned} \Sigma_{\Theta}^{fused} &= \left(\omega \left(\Sigma_{\Theta}^{L}\right)^{-1} + (1-\omega)\left(\Sigma_{\Theta}^{V}\right)^{-1}\right)^{-1} \\ \Theta^{fused} &= \Sigma_{\Theta}^{fused}\left(\omega \left(\Sigma_{\Theta}^{L}\right)^{-1}\Theta^{L} + (1-\omega)\left(\Sigma_{\Theta}^{V}\right)^{-1}\Theta^{V}\right) \end{aligned} \tag{4.14}$$

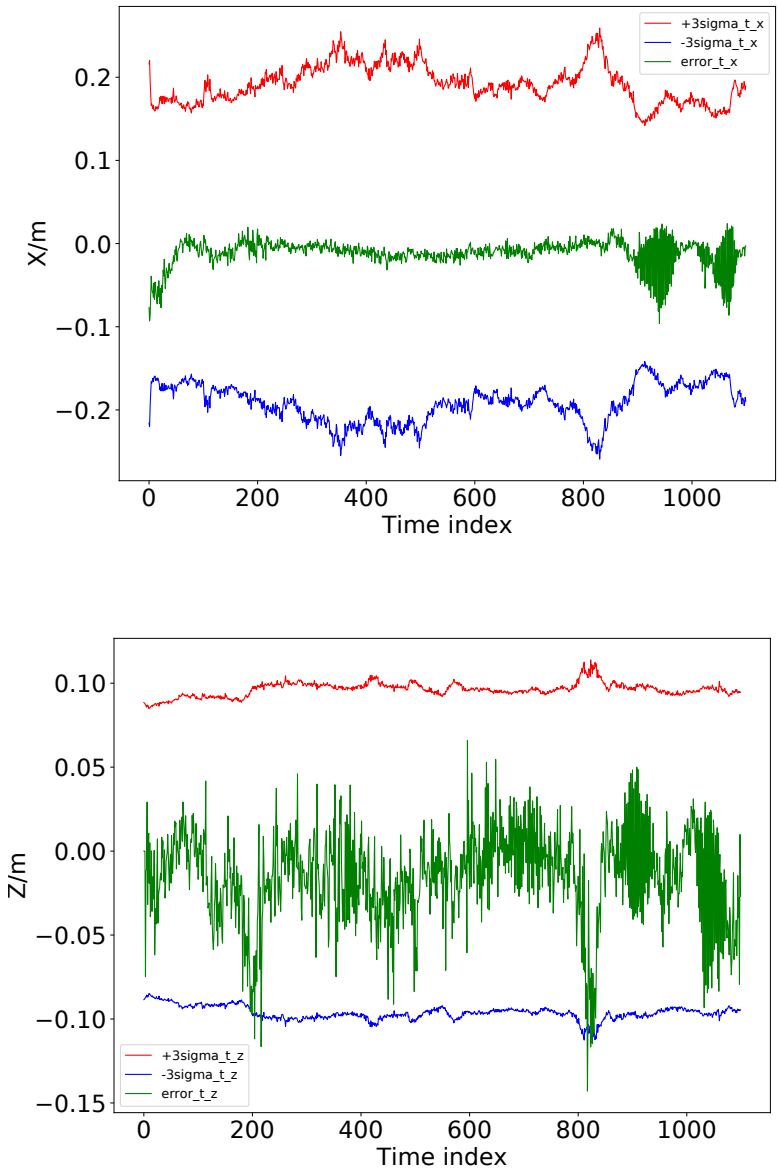


Figure 4.13 – Range sensor pose estimation uncertainty along t_x and t_z for KITTI sequence 01

where $\omega \in [0, 1]$ minimizes trace of the fused covariance matrix Σ_{Θ}^{fused} at each step. If the Jacobian matrix is near singular, probably because of local minimal occurrence or individual sensor failure, then inverting $\tilde{\mathbf{J}}_i'(\Theta^*)\Sigma_{\mathbf{x}_i}^{-1}\tilde{\mathbf{J}}_i(\Theta^*)$ will lead to unreliable uncertainty estimation marked as black dash lines in Fig. 4.12. Covariance intersection can ensure that the resulting estimate is conservative, which efficiently filters out the unstable

estimation. In order to simplify the fusion parameterization, only planar translation t_x and t_z and yaw angle r_y are fused and updated (see Fig. 4.14). The fused pose will better register the lidar map points and the multi-level voxel scan-to-map matching as in [23] is adopted to reduce the frame-to-frame estimation drift.

The KITTI dataset [55] contains stereo sequences and Velodyne HDL-64E LiDAR point clouds captured in urban and highway environments. The metric of average relative translation error t_{rel} proposed in [55] is used for evaluation purpose. The relative translation error t_{rel} measures the relative pose differences, which is suitable to evaluate the trajectories with various distances. Let ${}^{\mathcal{F}_W}\Theta_i$ and ${}^{\mathcal{F}_W}\hat{\Theta}_i$ be the reference and estimated ego-vehicle poses in the world frame, where i is the timestamp for each pose. The reference and estimated relative poses can be expressed as $\Delta_{i,j}$ and $\hat{\Delta}_{i,j}$, which are defined as follows:

$$\begin{aligned} \Delta_{i,j} &= ({}^{\mathcal{F}_W}\Theta_i)^{-1} \cdot {}^{\mathcal{F}_W}\Theta_j \\ \hat{\Delta}_{i,j} &= ({}^{\mathcal{F}_W}\hat{\Theta}_i)^{-1} \cdot {}^{\mathcal{F}_W}\hat{\Theta}_j, \\ t_{rel} &= \|(\Delta_{i,j})^{-1} \cdot \hat{\Delta}_{i,j}\|_2 \end{aligned} \tag{4.15}$$

To have a fair comparison, the ORB SLAM2 loop closure module is deactivated. Three typical sequences 01 (Highway), 02 (Urban+Country) and 07 (Urban) are chosen from KITTI dataset to make analysis and detailed quantitative result is shown in Tab. 4.3 [2]. The evaluation computes the relative translation errors for all possible subsequences of length and errors are measured in percent according to the average of those values. The proposed Loosely-Coupled Vision-LiDAR Odometry(LC-VLO) outperforms state-of-the art approaches for challenging trajectory in sequence 01. When driving on a highway scenario, few distinctive visual features are available (see Fig. 4.15), which makes descriptor-based feature tracking erroneous and thus causes poor pose estimation for visual sensor. Due to the uncertainty analysis and covariance intersection, the proposed approach ensures a consistent odometry estimation even in lack of usable visual features and moving at high speed (see Fig. 4.16(a)). In sequence 02, the proposed loosely coupled odometry is not as good as the ORB-SLAM2 due to the absence of horizontal lines or planes to constrain the drift along the vertical axis in the scan-to-map matching step. However, it does efficiently prevent large divergence occurrence like in A-LOAM [3] method. The large divergence mainly results from A-LOAM's inappropriate distance-based matching strategy. Far edge points are more likely to be mismatched when encountering large ro-

2. The sequence 08 is not evaluated due to ground truth flaw with manual inspection
3. Advanced implementation of LOAM, https://github.com/HKUST-Aerial-Robotics/A-LOAM

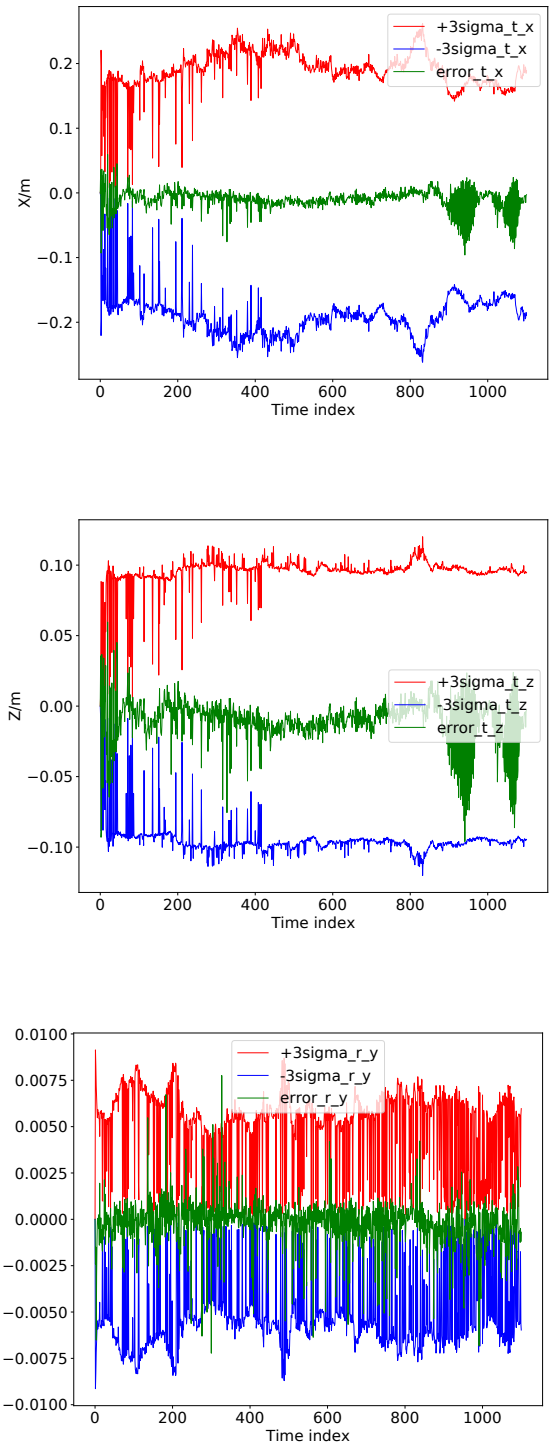


Figure 4.14 – Intersection fusion of planar translation t_x and t_z and yaw angle r_y

tational motion. It happens at the middle of sequence 02 where A-LOAM method loses the tracking of features and fails to confine the estimation error (see Fig. 4.16(b)). Uncertainty analysis is able to detect the potential deficiencies in the early scan-to-scan step and mitigate feature misalignment problem. The LC-VLO is superior to ORB-SLAM2 and A-LOAM for sequence 07, which shows that multi-level voxel scan-to-map matching procedure is indispensable to reduce frame-to-frame estimation drift (see Fig. 4.16(c)). Overall, the proposed LC-VLO adaptively fuses vision and LiDAR estimation, which is able to improve estimation performance for individual sensor degenerate cases, especially for the challenging KITTI sequence 01 and 02.

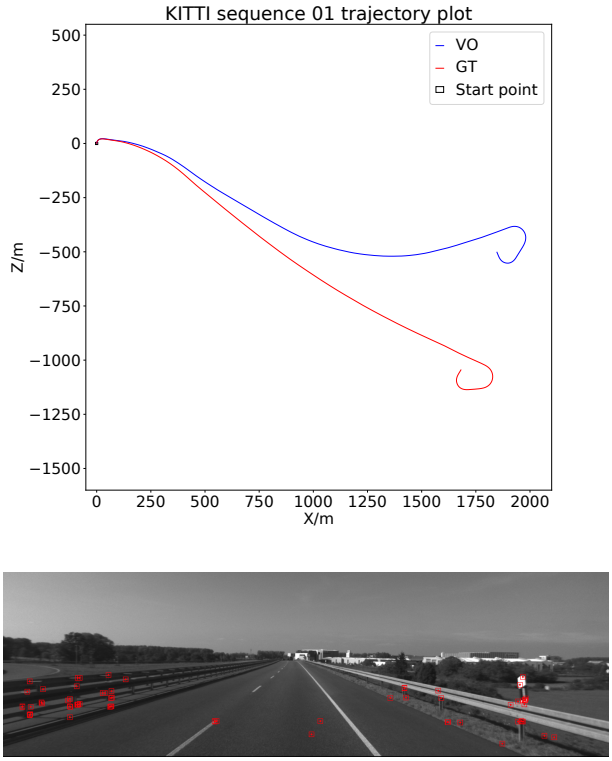
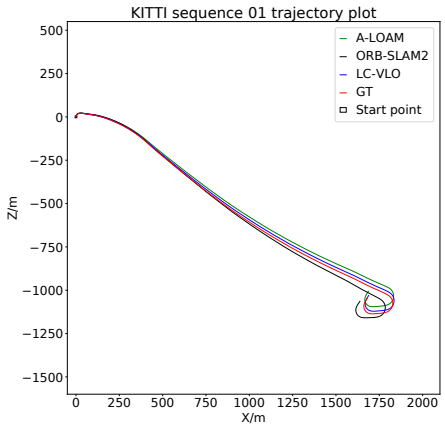


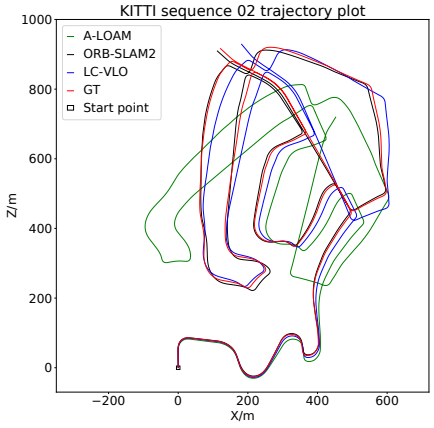
Figure 4.15 – KITTI sequence 01: Few distinctive ORB visual features for tracking on the highway scenario

4.4 Summary

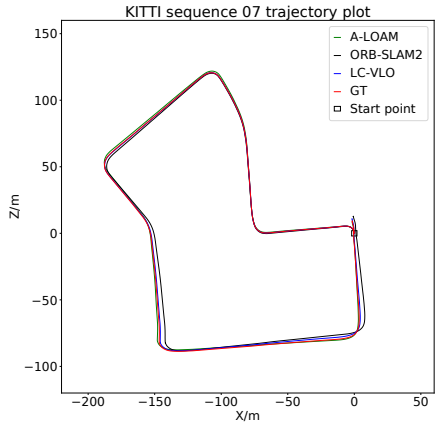
Reliable 3D perception is fundamental for localization, mapping and path-planning tasks of the autonomous driving. The proposed semantic-guided method fully exploits



(a)



(b)



(c)

Figure 4.16 – Estimated trajectory and ground-truth for KITTI 01, 02 and 07 sequences

Table 4.3 – Comparison of relative pose accuracy (%).

Metric t_{rel} / Sequence	ORB-SLAM2	A-LOAM	LC-VLO
00	0.88%	0.77%	**0.74**%
01	1.40%	2.27%	**0.84**%
02	**0.79**%	4.91%	1.50%
03	**0.77**%	1.24%	0.87%
04	**0.45**%	1.23%	1.08%
05	0.61%	0.70%	**0.43**%
06	0.73%	0.62%	**0.58**%
07	0.90%	0.63%	**0.50**%
08	–%	–%	–%
09	**0.90**%	1.09%	1.01%
10	0.59%	1.69%	**0.52**%

the range sensing capabilities of the multi-layer LiDAR for accurate objects detection, while levering the merits of image rich texture to determine state of motion. In this section, the effectiveness of the proposed approach to segment moving objects is highlighted by the comparison with the traditional robust kernel-based outlier rejection methods. The proposed approach is benchmarked with three city category sequences in the KITTI dataset, which outperforms the kernel-based methods and achieves the leading results of 77.9% average fitness and 7.65 cm RMSE respectively. Extensive qualitative and quantitative results demonstrate that, the proposed semantic-guided MOS helps to robustify the pose estimation process in challenging heavy traffic scenarios. Besides, the ghosting effect in scene reconstruction process is remarkably eliminated thanks to the moving objects removal.

Moreover, a loosely-coupled sensor fusion approach is developed, which efficiently combines complementary visual and range sensor information to estimate the vehicle ego-motion. Descriptor-based and distance-based matching strategies are respectively applied to visual and range measurements for feature tracking. Nonlinear optimization optimally estimates the relative pose across consecutive frames and an uncertainty analysis using forward and backward covariance propagation is made to model the estimation accuracy. Covariance intersection filter paves the way for us to loosely couple stereo vision and LiDAR odometry considering respective uncertainties. The proposed approach is evaluated with KITTI dataset which shows its effectiveness to fierce rotational motion and temporary absence of visual features, achieving the average relative translation error of 0.84% for

the challenging 01 sequence on the highway. This results from the anisotropic uncertainty modeling in the sensor fusion step, that does not misrepresent the potential errors. Since the sensor fusion is performed in a loosely-coupled manner, each sensor modality can be easily replaced according to personalized demands, which makes the proposed approach very flexible. As the current approach does not consider loop closure, more focus can be put on exploiting visual semantic hints for robust feature tracking and place recognition to further ameliorate the localization accuracy in the future.

Chapter 5

Conclusion and Perspectives

The reliable environment perception plays an important role in the autonomous driving system. As long as the intelligent vehicle has a good understanding of its surroundings, it can perform high-level tasks such as state estimation and motion planning under secure conditions. Hereby, state estimation incorporates both the ego-vehicle poses and map points location. This thesis concentrates on the moving objects detection and ego-motion estimation for the intelligent vehicles equipped with the camera and LiDAR sensors. In this Chapter, based on the obtained results in previous chapters, a synthesis of the thesis content will be given, which is followed by the future perspectives for the autonomous vehicle robust perception and state estimation module design.

5.1 Thesis Synthesis

To begin with, the vision-based perception pipeline is introduced in Chapter 2, where the mono-vision system is deployed to estimate the camera poses and reconstruct the sparse scene components jointly. It is demonstrated that the bundle adjustment embedded SfM approach exhibits good performance in indoor static environments. However, for the outdoor dynamic scenes, the presence of moving objects may pose a great challenge for the data association problem. For the state estimation task, the moving objects are considered as the noise, and should be filtered out beforehand. To address the issue of moving objects existence, the semantic-guided RANSAC approach is developed to associate the semantic information to the geometric entities. The moving objects are seamlessly identified with the combined epipolar and flow vector bound constraints. It is shown that the flow vector bound is indispensable to detect the moving objects which perform the degenerate motions. The semantic-guided RANSAC proves to be an efficient way to reject instance-level outliers, which meanwhile facilitates the estimation model convergence. In order to overcome the depth loss problem of mono-vision perception, the stereo-vision system estimates the pixel depth with the disparity and baseline distance. The 3D point

triangulation can be performed instantly with left-right image matching, which is more stable than the mono-vision triangulation with consecutive frames in aggressive motion cases. Given the 3D map points, the camera motion can be estimated with the 3D-2D reprojection residual error minimization. For the motion estimation uncertainty modeling, the backward covariance propagation is used to transform the uncertainty from the measurement domain to the estimation domain. The uncertainty quantification can reflect the estimation confidence and accuracy, on which basis, the sensor fusion can be implemented.

Then, the LiDAR sensor which is based on the laser reflection for the range measurement is used for environment perception and ego-motion estimation, in Chapter 3. Compared to visual sensors, LiDAR is invariant to the illumination conditions and its omnidirectional field of view enables scale-aware full pose estimation and precise 3D scene mapping. The relative transformation can be obtained by tightly aligning the consecutive dense point clouds. In Chapter 3, the investigations are conducted within the ICP framework for the laser scans alignment. Both direct and feature-based LiDAR scan matching approaches are benchmarked in various contexts, which includes parking, urban, and highway scenarios. In order to guarantee the scan matching performances in scenarios with scarce geometric information and fast ego-vehicle motion, an adaptive semi-direct scan matching method is proposed together with an alignment uncertainty quantification, which facilitates robust pose estimation and consistent scene reconstruction. To rectify the misaligned point clouds, the pose-graph optimization technique is applied to reduce the local drifts. A pose graph is formulated with edges representing pairwise registration, and then the deviated poses are penalized and smoothed with the global pose-graph optimization. The obtained results indicate that pose-graph optimization can significantly improve the pose estimation accuracy and scene mapping consistency, where the point clouds are well registered with high fidelity. Furthermore, the proposed semi-direct scan matching is tested on both the public KITTI and self-recorded LS2N datasets. It is demonstrated that this proposed approach outperforms the state-of-the-art and achieves the leading results with the RF and RMSE metrics. Additionally, for better understanding of the scene, the objects are segmented from the background with both model-based and learning-based methods. The detected 3D objects are considered as independent traffic participants and their 3D spatial location perception ensures the autonomous vehicle operation security in complex traffic scenes.

In the end, the complementary visual and range sensors are combined in Chapter 4, to

compensate individual sensing modality shortcomings. The sensor fusion and integration ensures the reliability of the autonomous vehicle perception system and facilitates the uncertainty-aware ego-motion state estimation. In Chapter 4, a semantic-guided LiDAR-vision fusion approach is proposed for efficient moving objects segmentation and robust ego-motion estimation. The proposed pipeline utilizes the LiDAR-based semantic segmentation as a prior and vision-based geometric information for validation. The effectiveness of this semantic-guided approach to segment moving objects is highlighted by the comparison with the traditional robust kernel-based outlier rejection methods. The approach is benchmarked with city category sequences in the KITTI dataset, which outperforms the kernel-based methods by a large margin. The robust ego-motion estimation is driven by the reliable correspondence matches across frames. And the existence of dynamic objects in the scene tends to cause the scan misalignment, thus degrading the registration accuracy of sequential LiDAR scans. Moving objects are considered as the most unstable traffic participants, which will corrupt the localization and mapping process. Since the moving objects are not temporally consistent, they do not belong to the permanent components of the scene. Therefore, moving objects should be eliminated from the mapping process in order to build a consistent representation of the scene. It proves that the proposed semantic-guided MOS helps to robustify the pose estimation process in challenging heavy traffic scenarios. Besides, the ghosting effect in scene reconstruction process is remarkably eliminated due to the moving objects removal. And the reconstructed static map will promote high-level tasks such as map-based localization and path planning. Afterwards, a loosely-coupled sensor fusion approach is developed in Section 4.3, which efficiently combines complementary visual and range sensor information to estimate the vehicle ego-motion. Descriptor-based and distance-based matching strategies are respectively applied to visual and range measurements for feature tracking. Nonlinear optimization optimally estimates the relative pose across consecutive frames and an uncertainty analysis using forward and backward covariance propagation is made to model the estimation accuracy. Uncertainty analysis is able to detect the potential deficiencies in the early scan-to-scan step and mitigate feature misalignment problem. Covariance intersection filter paves the way for us to loosely couple stereo vision and LiDAR odometry considering respective uncertainties. The proposed approach is evaluated with KITTI dataset, which shows its effectiveness to fierce rotational motion and temporary absence of visual features. This results from our anisotropic uncertainty modeling in the sensor fusion step, that does not misrepresent the potential errors. Since we perform the sensor fusion in a loosely-coupled

manner, each sensor modality can be easily replaced according to personalized demands, which makes our approach very flexible.

To conclude, the emerging autonomous vehicle will bring great transformation for the entire transportation industry. And for effective perception of the surroundings, multiple exteroceptive sensors need to be installed onboard. The different sensing modalities can capture different aspects of the world. Among them, the visual cameras are capable of recognizing the color and texture information of the scene. While, the LiDAR sensor is more robust for the range measurement collection. For the perception system, the multi-sensor measurement redundancy ensures the sensing reliability and accuracy. Meanwhile, it also requires efficient data processing and decision-making. Depending on the specific application scenarios and financial budgets, the sensor setup varies. However, the common issue to address before implementing the sensor fusion is to quantify the estimation uncertainty for individual sensor. In this thesis, accurate uncertainty modeling is achieved for both the vision-based and range-based estimation, which provides the possibility for adaptive sensor fusion. Moreover, when it comes to urban environments, pure geometric information is insufficient for the autonomous vehicle to have a comprehensive scene understanding. Thus, the semantic cues are exploited in this thesis for moving objects segmentation. The geometric and semantic information fusion helps to reject the outliers in the ego-motion estimation and scene mapping tasks, which also pushes the development of full autonomy intelligent vehicles.

5.2 Future Work

This thesis is dedicated to leveraging the camera and LiDAR sensors for the scene perception and ego-motion estimation. The LiDAR-vision fusion approach is explored in detail, where the semantic information and geometric constraints are combined for robust moving objects segmentation and ego-vehicle pose estimation. Despite the promising results that have been obtained, there are still some perspectives to make the improvement.

1. **Extrinsic Calibration.** In this thesis, we assume that the extrinsic parameters between the cameras and LiDAR sensors are well calibrated and remain unchanged. However, this assumption does not hold for long-term autonomous vehicle maneuvering. Therefore, regular extrinsic parameters calibration and synchronization are necessary when the sensor fusion system is on the run. And integrating the calibration pipeline into the perception system should be prioritized whenever the

multi-sensor measurements are fused.

2. **IMU Integration.** Inertial Measurement Unit (IMU) consists of a gyroscope to measure the angular velocity and an accelerometer to measure the linear acceleration. This type of inertial sensor estimates the ego-motion via the pre-integration [116] of angular velocity and linear acceleration, which is not affected by the illumination changes or the extreme weather condition. The inertial measurements can complement the state estimation whenever the feature tracking process fails. Additionally, the metric scale can be accurately recovered from the IMU measurement as long as the IMU bias is corrected. More efficient outliers rejection mechanisms that incorporate the IMU pre-integration can also be studied to further ameliorate the state estimation accuracy and robustness with the presence of several moving objects in the scene.
3. **Active Perception.** The ego-motion estimation in this thesis is restricted to passively extracting the environment information without any motion planning. While, active perception means that the autonomous vehicle can actively execute certain control commands, in order to reduce the estimation uncertainty. The control command is chosen in a predicted finite set by considering the motion constraints, in order to avoid cases such as the sudden decrease of inter-frame overlapping areas. And it is encouraged for the autonomous vehicle to drive to the previously visited area (loop closure), in order to correct the self-localization drift and to gain the confidence in the estimation results. At the same time, the semantic cues can also be used to identify the visited places, which boosts consistent state estimation and mapping.
4. **Mutual Communication.** In order to efficiently explore the large-scale unknown environment, multiple agents can collaborate to implement the perception task within each divided local area. The submaps can be built in the vicinity of each agent, and the local maps can be merged into a global one when the agents come across. The global pose graph can also be formed, with the nodes representing the agents' poses in each separated portion. The mutual communication can be achieved by the global consensus-based map representation and pose graph optimization.

BIBLIOGRAPHY

[1] Hans Peter Moravec, « Obstacle avoidance and navigation in the real world by a seeing robot rover », PhD thesis, Stanford University, 1980.

[2] Yang Cheng, Mark W Maimone, and Larry Matthies, « Visual odometry on the Mars exploration rovers-a tool to ensure accurate driving and science imaging », *in*: *IEEE Robotics & Automation Magazine* 13.*2* (2006), pp. 54–62.

[3] Mark Maimone, Yang Cheng, and Larry Matthies, « Two years of visual odometry on the mars exploration rovers », *in*: *Journal of Field Robotics* 24.*3* (2007), pp. 169–186.

[4] David Nistér, Oleg Naroditsky, and James Bergen, « Visual odometry », *in*: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 1, Ieee, 2004, pp. I–I.

[5] David Nistér, « An efficient solution to the five-point relative pose problem », *in*: *IEEE transactions on pattern analysis and machine intelligence* 26.*6* (2004), pp. 756–770.

[6] Martin A Fischler and Robert C Bolles, « Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography », *in*: *Communications of the ACM* 24.*6* (1981), pp. 381–395.

[7] Etienne Mouragnon et al., « Real time localization and 3d reconstruction », *in*: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, IEEE, 2006, pp. 363–370.

[8] Georg Klein and David Murray, « Parallel tracking and mapping for small AR workspaces », *in*: *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, IEEE, 2007, pp. 225–234.

[9] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos, « ORB-SLAM: a versatile and accurate monocular SLAM system », *in*: *IEEE transactions on robotics* 31.*5* (2015), pp. 1147–1163.

[10] Ethan Rublee et al., « ORB: An efficient alternative to SIFT or SURF », *in*: *2011 International conference on computer vision*, Ieee, 2011, pp. 2564–2571.

[11] Raul Mur-Artal and Juan D. Tardós, « Fast relocalisation and loop closing in keyframe-based SLAM », *in*: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 846–853, DOI: 10.1109/ICRA.2014.6906953.

[12] Raul Mur-Artal and Juan D Tardós, « Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras », *in*: *IEEE Transactions on Robotics* 33.*5* (2017), pp. 1255–1262.

[13] Sebastian Thrun, « Probabilistic robotics », *in*: *Communications of the ACM* 45.*3* (2002), pp. 52–57.

[14] Armin Hornung et al., « OctoMap: An efficient probabilistic 3D mapping framework based on octrees », *in*: *Autonomous robots* 34.*3* (2013), pp. 189–206.

[15] Ji Zhang and Sanjiv Singh, « LOAM: Lidar Odometry and Mapping in Real-time. », *in*: *Proceedings of Robotics: Science and Systems*, 2014, pp. 1–9.

[16] Wolfgang Hess et al., « Real-time loop closure in 2D LIDAR SLAM », *in*: *2016 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2016, pp. 1271–1278.

[17] Johannes Graeter, Alexander Wilczynski, and Martin Lauer, « Limo: Lidar-monocular visual odometry », *in*: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 7872–7879.

[18] Young-Sik Shin, Yeong Sang Park, and Ayoung Kim, « Direct visual SLAM using sparse depth for camera-lidar system », *in*: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2018, pp. 1–8.

[19] Xiao Liang et al., « Visual laser-SLAM in large-scale indoor environments », *in*: *Proceedings of the IEEE International Conference on Robotics and Biomimetics*, 2016, pp. 19–24.

[20] Zulun Zhu et al., « Loop detection and correction of 3d laser-based slam with visual information », *in*: *Proceedings of the 31st International Conference on Computer Animation and Social Agents*, 2018, pp. 53–58.

[21] Gaurav Pandey et al., « Visually bootstrapped generalized ICP », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2011, pp. 2660–2667.

[22] Ji Zhang and Sanjiv Singh, « Visual-lidar odometry and mapping: Low-drift, robust, and fast », *in*: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2015, pp. 2174–2181.

[23] Youngwoo Seo and Chih-Chung Chou, « A Tight Coupling of Vision-Lidar Measurements for an Effective Odometry », *in*: *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2019, pp. 1118–1123.

[24] Zhengyou Zhang, « A flexible new technique for camera calibration », *in*: *IEEE Transactions on pattern analysis and machine intelligence* 22.*11* (2000), pp. 1330–1334.

[25] Jakob Engel, Thomas Schöps, and Daniel Cremers, « LSD-SLAM: Large-scale direct monocular SLAM », *in*: *European conference on computer vision*, Springer, 2014, pp. 834–849.

[26] Jakob Engel, Jörg Stückler, and Daniel Cremers, « Large-scale direct SLAM with stereo cameras », *in*: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 1935–1942, DOI: 10.1109/IROS.2015.7353631.

[27] Chris Harris, Mike Stephens, et al., « A combined corner and edge detector », *in*: *Alvey vision conference*, vol. 15, 50, Citeseer, 1988, pp. 10–5244.

[28] J. Shi and C. Tomasi, « Good features to track », *in*: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600, DOI: 10.1109/CVPR.1994.323794.

[29] Edward Rosten and Tom Drummond, « Machine learning for high-speed corner detection », *in*: *European conference on computer vision*, Springer, 2006, pp. 430–443.

[30] David G Lowe, « Distinctive image features from scale-invariant keypoints », *in*: *International journal of computer vision* 60.*2* (2004), pp. 91–110.

[31] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, « Surf: Speeded up robust features », *in*: *European conference on computer vision*, Springer, 2006, pp. 404–417.

[32] Michael Calonder et al., « Brief: Binary robust independent elementary features », *in*: *European conference on computer vision*, Springer, 2010, pp. 778–792.

[33] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge University Press, 2004, DOI: 10.1017/CBO9780511811685.

[34] Richard I Hartley, « In defense of the eight-point algorithm », *in*: *IEEE Transactions on pattern analysis and machine intelligence* 19.6 (1997), pp. 580–593.

[35] Bill Triggs et al., « Bundle adjustment—a modern synthesis », *in*: *International workshop on vision algorithms*, Springer, 1999, pp. 298–372.

[36] Yi Ma et al., *An invitation to 3-d vision: from images to geometric models*, vol. 26, Springer, 2004.

[37] David G Lowe, « Object recognition from local scale-invariant features », *in*: *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, Ieee, 1999, pp. 1150–1157.

[38] Navneet Dalal and Bill Triggs, « Histograms of oriented gradients for human detection », *in*: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, Ieee, 2005, pp. 886–893.

[39] Ross Girshick et al., « Rich feature hierarchies for accurate object detection and semantic segmentation », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[40] Ross Girshick, « Fast r-cnn », *in*: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[41] Shaoqing Ren et al., « Faster r-cnn: Towards real-time object detection with region proposal networks », *in*: *Advances in neural information processing systems* 28 (2015).

[42] Joseph Redmon et al., « You only look once: Unified, real-time object detection », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[43] Joseph Redmon and Ali Farhadi, « YOLO9000: better, faster, stronger », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[44] Tsung-Yi Lin et al., « Microsoft coco: Common objects in context », *in*: *Proceedings of European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[45] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Second, Cambridge University Press, ISBN: 0521540518, 2004.

[46] Reza Sabzevari and Davide Scaramuzza, « Multi-body motion estimation from monocular vehicle-mounted cameras », *in*: *IEEE Transactions on Robotics* 32.*3* (2016), pp. 638–651.

[47] Bihao Wang, Sergio Alberto Rodríguez Florez, and Vincent Frémont, « Multiple obstacle detection and tracking using stereo vision: application and analysis », *in*: *Proceedings of International Conference on Control Automation Robotics & Vision (ICARCV)*, 2014, pp. 1074–1079.

[48] Dingfu Zhou et al., « Moving object detection and segmentation in urban environments from a moving platform », *in*: *Image and Vision Computing* 68 (2017), pp. 76–87.

[49] Ashit Talukder and Larry Matthies, « Real-time detection of moving objects from moving vehicles using dense stereo and optical flow », *in*: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2004, pp. 3718–3725.

[50] Abhijit Kundu, K Madhava Krishna, and Jayanthi Sivaswamy, « Moving object detection by multi-view geometric techniques from a single camera mounted robot », *in*: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009, pp. 4306–4312.

[51] Yanchao Yang et al., « Unsupervised moving object detection via contextual information separation », *in*: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 879–888.

[52] Katerina Fragkiadaki et al., « Learning to segment moving objects in videos », *in*: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4083–4090.

[53] Eric Brachmann and Carsten Rother, « Neural-guided RANSAC: Learning where to sample model hypotheses », *in*: *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4322–4331.

[54] Bruce D Lucas, Takeo Kanade, et al., « An iterative image registration technique with an application to stereo vision », *in*: *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981, pp. 24–28.

[55] Andreas Geiger et al., « Vision meets robotics: The KITTI dataset », *in*: *The International Journal of Robotics Research* 32.*11* (2013), pp. 1231–1237.

[56] G. Bradski, « The OpenCV Library », *in*: *Dr. Dobb's Journal of Software Tools* (2000).

[57] Kurt Konolige, « Small vision systems: Hardware and implementation », *in*: *Robotics research*, Springer, 1998, pp. 203–212.

[58] Heiko Hirschmuller, « Stereo processing by semiglobal matching and mutual information », *in*: *IEEE Transactions on pattern analysis and machine intelligence* 30.*2* (2007), pp. 328–341.

[59] P. J. Besl and N. D. McKay, « A method for registration of 3-D shapes », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.*2* (1992), pp. 239–256.

[60] Yang Chen and Gérard Medioni, « Object modelling by registration of multiple range images », *in*: *Image and Vision Computing* 10.*3* (1992), pp. 145–155.

[61] Sunil Arya and DM Mount, « Approximate Nearest Neighbor Queries in Fixed Dimensions », *in*: *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1993, pp. 271–280.

[62] D. Chetverikov et al., « The Trimmed Iterative Closest Point algorithm », *in*: *Proceedings of the IEEE International Conference on Pattern Recognition*, 2002, pp. 545–548, DOI: 10.1109/ICPR.2002.1047997.

[63] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun, « Generalized-icp. », *in*: *Proceedings of Robotics: science and systems*, 2009, pp. 435–443.

[64] Zhuping Wang et al., « Intelligent Vehicle Self-Localization Based on Double-Layer Features and Multilayer LIDAR », *in*: *IEEE Transactions on Intelligent Vehicles* 5.*4* (2020), pp. 616–625, DOI: 10.1109/TIV.2020.3003699.

[65] Pengwei Zhou et al., « T-LOAM: Truncated Least Squares LiDAR-Only Odometry and Mapping in Real Time », *in*: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–13, DOI: 10.1109/TGRS.2021.3083606.

[66] Jie Meng et al., « Efficient and Reliable LiDAR-Based Global Localization of Mobile Robots Using Multiscale/Resolution Maps », *in*: *IEEE Transactions on Instrumentation and Measurement* 70 (2021), pp. 1–15, DOI: 10.1109/TIM.2021.3093933.

[67] Radu Bogdan Rusu et al., « Aligning point cloud views using persistent feature histograms », *in*: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 3384–3391.

[68] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz, « Fast point feature histograms (FPFH) for 3D registration », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2009, pp. 3212–3217.

[69] Federico Tombari, Samuele Salti, and Luigi Di Stefano, « Unique signatures of histograms for local surface description », *in*: *Proceedings of the European Conference on Computer Vision*, 2010, pp. 356–369.

[70] François Pomerleau et al., « Comparing ICP variants on real-world data sets », *in*: *Autonomous Robots* 34.*3* (2013), pp. 133–148.

[71] Songming Chen and Vincent Frémont, « A Loosely Coupled Vision-LiDAR Odometry using Covariance Intersection Filtering », *in*: *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2021, pp. 1102–1107, DOI: 10.1109/IV48863.2021.9575275.

[72] Weikun Zhen et al., « A joint optimization approach of lidar-camera fusion for accurate dense 3-d reconstructions », *in*: *IEEE Robotics and Automation Letters* 4.*4* (2019), pp. 3585–3592.

[73] Mingyue Cui et al., « Dense Depth-Map Estimation Based on Fusion of Event Camera and Sparse LiDAR », *in*: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–11, DOI: 10.1109/TIM.2022.3144229.

[74] Tao Du et al., « An Integrated INS/Lidar Odometry/Polarized Camera Pose Estimation via Factor Graph Optimization for Sparse Environment », *in*: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–11, DOI: 10.1109/TIM.2022.3156976.

[75] Andy Zeng et al., « 3dmatch: Learning local geometric descriptors from rgb-d reconstructions », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1802–1811.

[76] Haowen Deng, Tolga Birdal, and Slobodan Ilic, « Ppfnet: Global context aware local features for robust 3d point matching », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 195–205.

[77] Zi Jian Yew and Gim Hee Lee, « 3dfeat-net: Weakly supervised local 3d features for point cloud registration », *in*: *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 607–623.

[78] Michelle Valente, Cyril Joly, and Arnaud de La Fortelle, « An LSTM network for real-time odometry estimation », *in*: *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2019, pp. 1434–1440.

[79] G Dias Pais et al., « 3dregnet: A deep neural network for 3d point registration », *in*: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7193–7203.

[80] Xieyuanli Chen et al., « OverlapNet: Loop closing for LiDAR-based SLAM », *in*: *arXiv preprint arXiv:2105.11344* (2021).

[81] Christian Forster, Matia Pizzoli, and Davide Scaramuzza, « SVO: Fast semi-direct monocular visual odometry », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2014, pp. 15–22.

[82] Xiangrui Zhao et al., « A robust stereo feature-aided semi-direct SLAM system », *in*: *Robotics and Autonomous Systems* 132 (2020), p. 103597.

[83] Nicola Krombach, David Droeschel, and Sven Behnke, « Combining feature-based and direct methods for semi-dense real-time stereo visual odometry », *in*: *Proceedings of the International Conference on Intelligent Autonomous Systems*, 2016, pp. 855–868.

[84] Yu Zhong, « Intrinsic shape signatures: A shape descriptor for 3d object recognition », *in*: *IEEE International Conference on Computer Vision Workshops*, 2009, pp. 689–696.

[85] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun, « Open3D: A Modern Library for 3D Data Processing », *in*: *arXiv:1801.09847* (2018).

[86] Andrea Censi, « An accurate closed-form estimate of ICP's covariance », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2007, pp. 3167–3172, DOI: 10.1109/ROBOT.2007.363961.

[87] Anders Glent Buch, Dirk Kraft, et al., « Prediction of ICP pose uncertainties using Monte Carlo simulation with synthetic depth images », *in*: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 4640–4647.

[88] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun, « Robust reconstruction of indoor scenes », *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5556–5565.

[89] Rainer Kümmerle et al., « G2o: A general framework for graph optimization », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2011, pp. 3607–3613, DOI: 10.1109/ICRA.2011.5979949.

[90] Martin Ester et al., « A density-based algorithm for discovering clusters in large spatial databases with noise. », *in*: *kdd*, vol. 96, 34, 1996, pp. 226–231.

[91] John A Hartigan and Manchek A Wong, « Algorithm AS 136: A k-means clustering algorithm », *in*: *Journal of the royal statistical society. series c (applied statistics)* 28.*1* (1979), pp. 100–108.

[92] Charles R Qi et al., « Pointnet: Deep learning on point sets for 3d classification and segmentation », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[93] Yin Zhou and Oncel Tuzel, « Voxelnet: End-to-end learning for point cloud based 3d object detection », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[94] Alex H Lang et al., « Pointpillars: Fast encoders for object detection from point clouds », *in*: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12697–12705.

[95] Wei Liu et al., « Ssd: Single shot multibox detector », *in*: *European conference on computer vision*, Springer, 2016, pp. 21–37.

[96] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl, « Center-based 3d object detection and tracking », *in*: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11784–11793.

[97] Martin Simon et al., « Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds », *in*: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[98] Vincent Frémont, Sergio Alberto Rodriguez Florez, and Bihao Wang, « Monocular vision based moving object detection in complex traffic scenes », *in*: *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2017, pp. 1078–1084.

[99] Youbing Wang and Shoudong Huang, « Towards dense moving object segmentation based robust dense RGB-D SLAM in dynamic scenarios », *in*: *Proceedings of the IEEE International Conference on Control Automation Robotics and Vision*, 2014, pp. 1841–1846.

[100] Rahul Kumar Namdev et al., « Motion segmentation of multiple objects from a freely moving monocular camera », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2012, pp. 4092–4099.

[101] Songming Chen, Haixin Sun, and Vincent Frémont, « Mono-Vision based Moving Object Detection using Semantic-Guided RANSAC », *in*: *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2021, pp. 1–6, DOI: 10.1109/MFI52462.2021.9591165.

[102] Bihao Wang, Sergio Alberto Rodríguez Florez, and Vincent Frémont, « Multiple obstacle detection and tracking using stereo vision: Application and analysis », *in*: *Proceedings of the IEEE International Conference on Control Automation Robotics and Vision*, 2014, pp. 1074–1079, DOI: 10.1109/ICARCV.2014.7064455.

[103] Sergio A. Rodríguez F et al., « Visual confirmation of mobile objects tracked by a multi-layer lidar », *in*: *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2010, pp. 849–854, DOI: 10.1109/ITSC.2010.5625200.

[104] Philippe Babin, Philippe Giguere, and François Pomerleau, « Analysis of robust functions for registration algorithms », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2019, pp. 1451–1457.

[105] Philipp Ruchti and Wolfram Burgard, « Mapping with Dynamic-Object Probabilities Calculated from Single 3D Range Scans », *in*: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 6331–6336, DOI: 10.1109/ICRA.2018.8463149.

[106] Hanyu Shi et al., « SpSequenceNet: Semantic Segmentation Network on 4D Point Clouds », *in*: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4573–4582, DOI: 10.1109/CVPR42600.2020.00463.

[107] Giseop Kim and Ayoung Kim, « Remove, then Revert: Static Point cloud Map Construction using Multiresolution Range Images », *in*: *Proceedings of the IEEE/RSJ*

International Conference on Intelligent Robots and Systems, 2020, pp. 10758–10765, DOI: 10.1109/IROS45743.2020.9340856.

[108] Xieyuanli Chen et al., « Moving object segmentation in 3D LiDAR data: A learning-based approach exploiting sequential data », *in*: *IEEE Robotics and Automation Letters* 6.*4* (2021), pp. 6529–6536.

[109] Chen Fu et al., « Camera-Based Semantic Enhanced Vehicle Segmentation for Planar LIDAR », *in*: *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2018, pp. 3805–3810, DOI: 10.1109/ITSC.2018.8569413.

[110] Khaled El Madawi et al., « Rgb and lidar fusion based 3d semantic segmentation for autonomous driving », *in*: *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 2019, pp. 7–12.

[111] Rafael Barea et al., « Vehicle Detection and Localization using 3D LIDAR Point Cloud and Image Semantic Segmentation », *in*: *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2018, pp. 3481–3486, DOI: 10.1109/ITSC.2018.8569962.

[112] Abhijit Kundu, K Madhava Krishna, and Jayanthi Sivaswamy, « Moving object detection by multi-view geometric techniques from a single camera mounted robot », *in*: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 4306–4312, DOI: 10.1109/IROS.2009.5354227.

[113] Dorian Galvez-López and Juan D. Tardos, « Bags of Binary Words for Fast Place Recognition in Image Sequences », *in*: *IEEE Transactions on Robotics* 28.*5* (2012), pp. 1188–1197, DOI: 10.1109/TRO.2012.2197158.

[114] Simon J Julier and Jeffrey K Uhlmann, « Using covariance intersection for SLAM », *in*: *Robotics and Autonomous Systems* 55.*1* (2007), pp. 3–20.

[115] Edward H Adelson et al., « Pyramid methods in image processing », *in*: *RCA engineer* 29.*6* (1984), pp. 33–41.

[116] Todd Lupton and Salah Sukkarieh, « Visual-Inertial-Aided Navigation for High-Dynamic Motion in Built Environments Without Initial Conditions », *in*: *IEEE Transactions on Robotics* 28.*1* (2012), pp. 61–76, DOI: 10.1109/TRO.2011.2170332.

Titre : Perception et estimation d'état basées sur plusieurs capteurs pour les véhicules autonomes

Mot clés : Perception de l'environnement, Estimation d'état, Analyse d'incertitude, Fusion de capteurs, Véhicules autonomes

Résumé : Percevoir ou comprendre les environnements environnants est indispensable pour construire des systèmes d'aide à la conduite ou des véhicules autonomes. Dans cette thèse, nous étudions l'approche de fusion de capteurs pour le problème de localisation et de cartographie simultanées (SLAM) avec des capteurs visuels et de distance complémentaires. Afin de prendre des décisions conservatrices et d'augmenter la sécurité de manœuvre des véhicules autonomes, l'analyse d'incertitude de l'estimation de la pose est également mise en œuvre. Le système SLAM traditionnel suppose des scènes statiques, ce qui est vulnérable dans le contexte d'environnements extérieurs dynamiques. Ainsi, nous introduisons une approche basée sur les données pour exploiter les informations sémantiques qui interprètent la séquence de mesure à travers les cadres, ce qui distingue efficacement les objets en mouvement des objets statiques. Nous testons les algorithmes proposé sur des données réelles de trafic urbain et d'aires de stationnement, qui présente des résultats prometteurs.

Title: Multi-Sensor based Perception and State Estimation for Autonomous Vehicles

Keywords: Environment perception, State estimation, Uncertainty analysis, Sensor fusion, Autonomous vehicles

Abstract: Perceiving or understanding the surrounding environments is indispensable for building driving assistant systems or autonomous vehicles. This thesis studies the sensor fusion approach for the simultaneous localization and mapping (SLAM) problem with complementary visual and range sensors. In order to make conservative decisions and increase autonomous vehicle maneuvering security, the uncertainty analysis of the pose estimation is also implemented. The traditional SLAM system has the assumption of static scenes, which is vulnerable in the context of dynamic outdoor environments. Thus, we introduce a data-driven approach to exploit the semantic information that interprets the measurement sequence across the frames, which efficiently distinguishes the moving objects from the static ones. We test the proposed algorithms on real-life data of urban traffic and parking areas, which presents promising results.