Vision-Based Localization and Map Update for Long-Term Navigation in Changing Environments

by Jingwen YU
Supervisor: Prof. Ming LIU

Department of Electronic and Computer Engineering

The Hong Kong University of Science and Technology

Abstract

For autonomous robot navigation, a vision-based system can be easily deployed with inexpensive commercial cameras, while visual sensors are sensitive to environmental changes caused by lighting, object movement, weather, and seasons. For both indoor and outdoor autonomous navigation, the changes over time lead to poor localization and deterioration in map quality. This may result in a significant drop in robustness, especially for tasks involving repeated traversals such as autonomous path following, i.e. visual teach-and-repeat task. To enable a long-term autonomous navigation system, the robustness of the system is a must. While extensive techniques are proposed to strengthen the robustness of localization through image retrieval or deep-trained image features that are invariant to appearance changes and viewing angle. They rarely focus on the changes caused by object movement. These kinds of changes are rather typical in situations such as parking lots, warehouses, and offices, which cause not only the appearance change but also the infeasibility of the previous paths. In extreme conditions, object movement may lead to an insufficient number of feature points. Besides, localization and mapping are tightly-coupled, however, current research lacks an emphasis on map management and update for long-term navigation tasks. Therefore, this proposal focuses on the long-term navigation system enabled by visual sensors and targets to test and improve localization robustness under environmental changes aroused by object movements. Moreover, the corresponding map maintenance algorithms aiming at recording and updating the surrounding changes are proposed as an important research direction. With the robustness improvement on localization and mapping under changing environments, the long-term autonomous navigation tasks e.g. autonomous route following should be achieved.

CHAPTER 1

INTRODUCTION

1.1 Background

Autonomous mobile robotic systems are emerging dramatically in recent years from unmanned aerial vehicles (UAVs) to unmanned ground vehicles (UGVs), from self-driving vehicle to indoor service robots. A large portion of industrial applications has already achieved high-level autonomy through state-of-the-art techniques. For industrial robots, the operation environment is usually structurally constrained and fixed with various customized assisting modules such as using a pre-installed external visual marker system to assist localization and detection. Therefore, the industrial environment is barely changed with high stability, while for other demanding scopes of applications (e.g. autonomous driving, warehouse robots, and service robots), the operational surrounding changes significantly with high dynamics and uncertainties. This highlights the key focus on robustness and reliability under changing environments for the wide spread of autonomous mobile robots. In the development of Hercules[27], an autonomous vehicle targeting low-speed delivery, the stability under the dynamic environment is demonstrated and emphasized to enable safe autonomy. A picture of Hercules is shown in Fig4.1.

Compared to a 3D laser scanner, also known as Light detection and ranging (LiDAR), visual sensors (i.e. cameras) are cheaper and able to provide rich texture information, which is more suitable for high-level interactive tasks such as mobile manipulation, and human-robot interaction. Cameras are also lightweight which enables massive and efficient deployment on all kinds of robotic platforms, e.g. on micro aerial robotics. While LiDARs are not suitable for their weight and size, as well as high drive power source requirements. Therefore, vision-based localization and mapping techniques are important which can achieve comparable performance with much lower hardware requirements.

This proposal focuses on the vision-based long-term navigation tasks, which require high robustness. During the task completion, the robot might encounter an environment with uncertainty, such as highly dynamic objects (humans, vehicles, and other operating robots), moderately dynamic objects (boxes, bins, and chairs), and static objects but

could be changed or moved (fences, wall, and trees). This uncertainty of surrounding changes can lead to the failure of localization and inaccuracy of the map. Without a doubt, the localization of robots is the fundamental problem that other operations rely on. Regarding map accuracy, the incremental changes and possible loss of localization make it important to maintain an up-to-date map. Moreover, the discovery of the human memory system shows the mechanism behind an incrementally and continuous learning system. The human memory system contains temporal memory as well as long-term memory. In 2022, BioSLAM[51] proposed a dual-memory system for place recognition system to maintain accuracy with various incremental new appearances. Current research seldom tackles this problem while mostly focusing on temporal navigation and assuming the environment to be static. Constrained by real-time computing resources as well as robustness, there is still a large gap toward long-term stable mobile navigation. Inspired by [51] and [33], both adopted the idea of building more than one map over the same changing environment, I proposed to carry out further research on strengthening the robustness of long-term perception system by leveraging multiple experiences and traversals over the same place. Specifically, from two perspectives, localization algorithm and map management mechanism.

1.2 Related Work

1.2.1 Visual Teach-and-Repeat

The teach-and-repeat navigation[12, 29] is one of the basic tasks in mobile robots, which enables a large number of applications such as autonomous delivery, transportation, inspection, and patrolling. As the name suggests, the system contains two phases. In the teaching phase, the robot usually travels along the desired trajectory by manual operation. And in the repeat phase, the autonomously follows the previously taught trajectory. This navigation system can be developed based on a variety of sensors including cameras[12, 29, 9, 8], LiDARs[41], GPS[26], wheel encoders[24] and sensor fusion[31]. In this proposal, the adopted sensors are limited to visual sensors, including monocular cameras, stereo cameras, and RGB-D cameras. Regarding the visual teach and repeat system, [9] achieves fast and robust teach and repeat with a topological map containing a sequence of ordered images. Furgale and Barfoot[12] combine topological and metric map representation to build a manifold of overlapping locally consistent submaps. provides a state-of-the-art algorithm that enables navigation leveraging a topo-metric

map. Paton[33] further extends the system to record multiple repeated experiences with a Spatio-temporal pose graph to mitigate the incremental appearance changes. Most repeated tasks (e.g. inspection, delivery, and patrolling), are expected to be continuously repeated with a single teaching process, which is a typical scenario that the appearance changing would be challenging over a long time[34]. Due to the extra uncertainty brought by the changes, it is difficult to achieve robust navigation with only an initial static map. Furthermore, this uncertainty in observation is also vital in various navigation tasks, such as a self-driving system that makes use of a high-definition map to implement localization and path planning. The changes in road lanes can't be ignored. To this end, this proposal was originally inspired by the observation of the teach-and-repeat system and expected to extend the solution to general mobile robot systems.

1.2.2 Long-term Metric Localization

Accurate localization is an essential problem for autonomous robot operations. Vision-based localization is susceptible to appearance changes primarily due to the variation of spare image descriptors (e.g. SURF[2]) under changing environments. An intuitive solution is to design an appearance invariant local feature. Learning-based methods have achieved impressive performance[10, 35, 44, 14] over lighting changes of day and night, weather, and seasons compared to handcraft features[37, 2]. In the meantime, visual place recognition (VPR) techniques have been developed for the image retrieval task, where the query image location is estimated using the localizations of the most visually similar images obtained from an image database. This task plays an important role in loop closure detection and re-localization problem. Convolutional Neural Network (CNN) based methods, led by NetVLAD[1] achieve recognition under large-scale, extreme viewpoint and appearance variation conditions. VPR provides a global topological localization, while in most navigation tasks, the six-degree-of-freedom (6DoF) metric localization is required to perform complex tasks. The state-of-art metric localization methods combine the VPR with sparse feature matching[38, 13, 39], which is known as hierarchical localization. This localization framework leads the state-of-the-art performance on the long-term visual localization benchmark. As the deep learning techniques enable semantics inference, which promotes the research on leveraging the semantic property of the scene to enhance the localization[25, 42, 45, 54, 30]. These works proposed to deal with long-term visual localization based on the semantics of the scene that appears to be invariant. This is true in most scenarios where the appearance changes are induced by the lighting

variation, different weather conditions, and seasons. However, the changes caused by the movement of the object are not issued or evaluated in these works. In extreme situations, the movement of large-scale objects (e.g. buildings, fences, and trees) may cause the occlusion or newly appeared scenes. Long-term metric visual localization remains a challenge under dynamic conditions which has not been investigated thoroughly.

1.2.3 Mapping Under Dynamic Scenes

Accurate localization serves as the essential preliminary for globally consistent maps. There are different map representations or data structures designed for different purposes. Kimera[36] provides a comprehensive collection of layered map representations, which reconstruct the environment from bottom to top. From the bottom of the metric point cloud layer to the top topological layer, different layers serve different robotic applications. Despite the various map representations constructed by Kimera, it does not highlight the uncertainty caused by dynamic objects. Within an environment, the objects can be classified as dynamic objects, semi-static objects, and static objects. [22] proposed direct dynamic object removal on the point cloud, and further extend to a modular SLAM framework[23]. This algorithm is suitable for continuously moving objects within the field of view, while it could not remove the objects that are moving under constant speed with respect to the sensor's ego-motion. Works have been done on dealing with dynamic objects and building dense volumetric maps based on truncated signed distance functions (TSDFs) [40, 15, 11, 32, 34]. POCD[34] and Panoptic Mapping[40]put focus on the indoor semi-static environments. For example, the furniture movement, warehouse goods movement, and parked cars movement in the parking lot. They achieve improvement over map maintenance, aiming to capture the environmental changes accurately and efficiently with the assumption of perfect localization is already provided. The tests on mapping under inaccurate or corrupted pose estimation have not been done to demonstrate the robustness of such a mapping system. Visual localization under changing environments remains an open problem, especially in the background of teach-and-repeat tasks. The map update system depends on the successful completion of repeated traversals. It is also more reasonable and practical to discuss the mapping under dynamic scenes combined with uncertainty localization primitive conditions.

CHAPTER 2

PROPOSED DIRECTION

Based on the above investigations over long-term localization[38, 33] and mapping[34, 40] algorithms, the proposed directions lie in following three parts. For current map maintenance systems, accurate 6DoF localization is assumed to be solved and provided as a precondition. Therefore, further tests and research on the localization in dynamic scenes are proposed. Regarding the mapping system, multi-experiences localization (MEL)[33] uses multiple traversals information to continuously build topo-metric maps. It proposed a lifelong learning mapping system without considering the large-scale problem, the map size is not scalable over long-term operation. While POCD[34] maintains the map consistent online without considering that the lighting conditions might be changed significantly. Maintenance of multiple maps of the same scene by typical appearance conditions (e.g. day and night, summer and winter) may improve the robustness of long-term navigation.

2.1 Localization in Dynamic Scenes

Accurate localization guarantees strong preliminary conditions for mapping. In dynamic scenes, the localization might be difficult due to the occlusions. Recent works on Simultaneous Localization and Mapping (SLAM) [53, 48, 17] have issued this problem and proposed to remove the dynamic objects via object detection or semantic segmentation algorithms. Excluding the feature points extracted from dynamic objects provides static feature points for visual odometry to track location. It is possible to adopt this kind of idea and integrate it into HLoc [38] which achieves state-of-the-art performance in localization under an appearance-changing environment. From another standing point,[15] using TSDFs to generate a dynamic-object aware scene reconstruction with multiple dynamic objects. It can withstand occlusions introduced by object movement, which is a promising scene representation suitable for dynamic scenes. Therefore, localization in a TSDF-based map can be an interesting direction to dig in. Existing works proposing localization in dense scene reconstruction (3D surfel map[50], signed distance field

[20]) proved the possibility of such an approach. And the last proposed direction is to investigate further research via experience-based navigation. [33]proposed a localization method to localize across different traversals. Further inspired by [51], an online learning mechanism can make use of multiple experiences to improve localization performance. MEL could be combined with current data-driven approaches to serve as an optimization strategy for online learning of long-term localization.

2.2 Map Update and Management

POCD[34] and Panoptic Mapping[40] show contributions to the maintenance of online map consistency, which serves as an excellent improvement for robot environment interaction tasks (e.g. mobile manipulation). Based on the experiment result demonstrated by MEL[33], using multiple maps of different environmental conditions makes a significant boost to the localization methods. In the meantime, TSDF-based methods[11, 15] prove truncated signed distance functions to be an effective map representation under changing environments. Therefore, the construction and maintenance of multiple TSDF-based maps could be a potential research direction. As brought by MEL, the scalability of the map is an open problem. In the MEL system, there is no management over experiences, which means all the repeated traversals are stored, which could lead to the large data stored and the exhaustive search over increasingly larger Spatio-temporal-pose-graph could lead to inefficiency over long-term operation. There are two possible approaches to this scalability issue. One is to use a limited number of experiences, which involve the inter-experience evaluation, with designed criteria such as using semantics to guide the metrics over inter-experience changes. An intuitive idea for a cross-season map is to filter out one representative experience for each season. While POCD classifies the environment into dynamic objects, semi-static objects, and static objects, using semantics can be strong prior to potential changes that may appear on the map. However, scene semantic inference remains an open problem since current semantic segmentation and panoptic segmentation algorithms are limited by domain adaptation, generalization, and extensive training data demands. Another direction that remains open is the data structure of the map, there has been a focus in the LiDAR-based SLAM system (e.g. iKD-Tree[4], Octree[18]). Similar approaches could be adopted to the visual community with incrementally increased spare feature points as well as potential properties (e.g. semantics, weather, seasons).

2.3 Dataset

Dataset is essential in order to research long-term visual localization and mapping. Because current popular datasets such as EUROC[3], TUM[43], and ICL-NUIM[16], mostly focus on localization and mapping within one traversals. The repeated observations are in a short period, which is limited to testing the long-term localization and mapping performance. For VPR, there are Oxford Robocar[28], NCLT[5], and Pittsburgh250k[46]. The main focus is the illumination change and viewpoint change. While VPR datasets provide image sequences over different times of the day or even across seasons, they do not focus on structure changes over time which may happen indoors. TorWICD[34] and RIO[47] provides a good example of object level changes under multiple repeated traversals. They lack 3D laser scan data which could be used to achieve accurate pose estimation and 3D scene reconstruction. In order to generate a benchmark for long-term visual localization and mapping, I proposed to use a portable sensor fusion platform (FusionPortable) that could be deployed to a wide range of mobile platforms (e.g. wheel robots, quadruped robots, and hand-held devices). Using the FusionPortable platform, monocular camera, stereo camera, and event camera as well as 3D laser scenario, IMU, GPS, and RTK data could be acquired. With mobile platform quadruped robot, autonomous vehicle, and hand-held device, the dataset could provide cross-platform, long-term, object-level aware data under dynamic and static environments on campus. To get accurate structural change caused by semi-static object movement, we proposed to use high definition image laser scanner (i.e. Leica BLK360 G1) to get the high-resolution 3D reconstruction of the indoor office over a long period.

CHAPTER 3

PRELIMINARY WORK

In this chapter, preliminary works toward the proposed directions have been conducted which also serve as the root of motivation and inspiration.

3.1 Localization and Mapping

In this work, I am trying to build an indoor mapping and localization system for a quadruped robot to enable navigation tasks. For the mapping part, LiDAR odometry and mapping (LOAM, A-LOAM[55]) are used during the localization part. In order to implement obstacle avoidance and path planning, an occupancy map is implemented online. However, when an object exists in an occupancy map for a long time, even after it disappears, the occupancy grid occupied by it won't be released. This is due to its bayesian update nature which brings map update to my attention. Part of the result is shown in Fig4.2 and Fig 4.3.

3.2 Semantic Segmentation

In this project, I got hands-on experience in training a semantic segmentation network from scratch and via fine-tuning, which achieves the segmentation tasks on a private and small-scale autonomous driving dataset with self-defined semantic classes. This serves as an evaluation of the state-of-the-art segmentation networks' generalization and domain adaptation abilities. In this project, DeepLabV3[6], DeepLabV3+[7], MobileNetV3[19], and BiSeNetV2[52] have been tested with results in Fig4.4 .

3.3 2D Object Detection

In this project, a relationship-oriented perception pipeline for accomplishing daily manipulation tasks is implemented. 2D object detection and semantic relationship inference are combined and explored to achieve object-level manipulation tasks. This

builds up my experience in 2D object detection and inspired me to further research the semantic relationship between objects within a certain scene to construct high-level constraints over the consistency of map semantics. Partial results are shown in Fig4.5. This work has been accepted by the 2022IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS2022) with more information can be referred to `https://www.youtube.com/watch?v=48rpq9SoPYo`.

3.4 Visual Place Recognition

Visual place recognition (VPR) in condition-varying environments is still an open problem. In this paper[49], we propose to use a convolutional autoencoder (CAE) to tackle this problem. We employ a high-level layer of a pre-trained CNN to generate features and train a CAE to map the features to a low-dimensional space to improve the condition invariance property of the descriptor and reduce its dimension at the same time (detailed approach shown in Fig4.6). We verify our method in three challenging datasets involving significant illumination changes, and our method is shown to be superior to the state-of-the-art. The code is publicly available in `https://github.com/MedlarTea/CAE-VPR`.

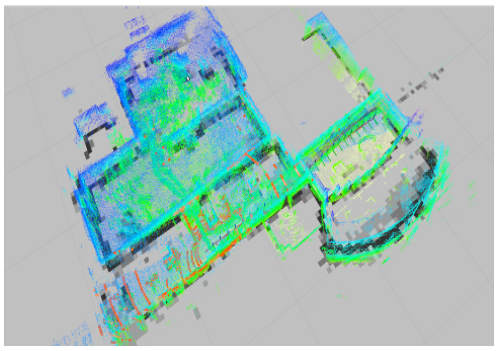
3.5 Multi-Sensor Fusion Dataset

This paper[21] proposes the FusionPortable benchmark, a novel multi-sensor dataset with a diverse set of sequences for mobile robots, shown in Fig . This is a great fit for the collection of real-world long-term datasets mentioned in Section 2.3. We first advance a portable and versatile multi-sensor suite that offers rich sensory information: 10Hz LiDAR point clouds, 20Hz stereo frame images, high-rate and asynchronous events from stereo event cameras, 200Hz acceleration and angular velocity readings from a tactical grade IMU, and 10Hz GPS signal outdoors. Sensors are already temporally synchronized in hardware. This device is lightweight, self-contained, and has plug-and-play support for mobile robots. Second, we construct a dataset by collecting 18 sequences that cover a variety of environments on the campus by exploiting multiple platforms for data collection.

CHAPTER 4

LIST OF FIGURES

Figure 4.1: The first autonomous vehicle (AV) trial without an operator on board in Hong Kong commenced in late 2020 on the HKUST Clearwater Bay Campus. The AV, designed by Prof. Ming Liu (ECE, Director of HKUST's Intelligent Autonomous Driving Centre) and his team of students, is one of the many innovative initiatives to fight against the COVID-19 pandemic as the AV can make deliveries that limit human-to-human contact.



(a) A point cloud map of the second floor, HKUST CYT constructed by A-LOAM

(b) A capture of localization on quadruped robot in an indoor office environment.

Figure 4.2: Mapping and Localization of Quadruped Robot

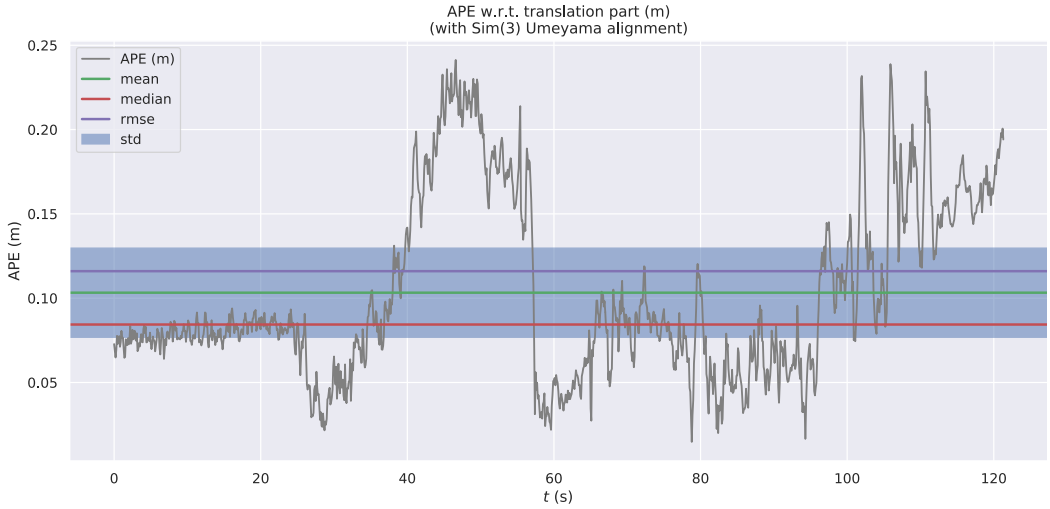


Figure 4.3: Qualitity evaluation result of localization accuracy in the indoor office environment.

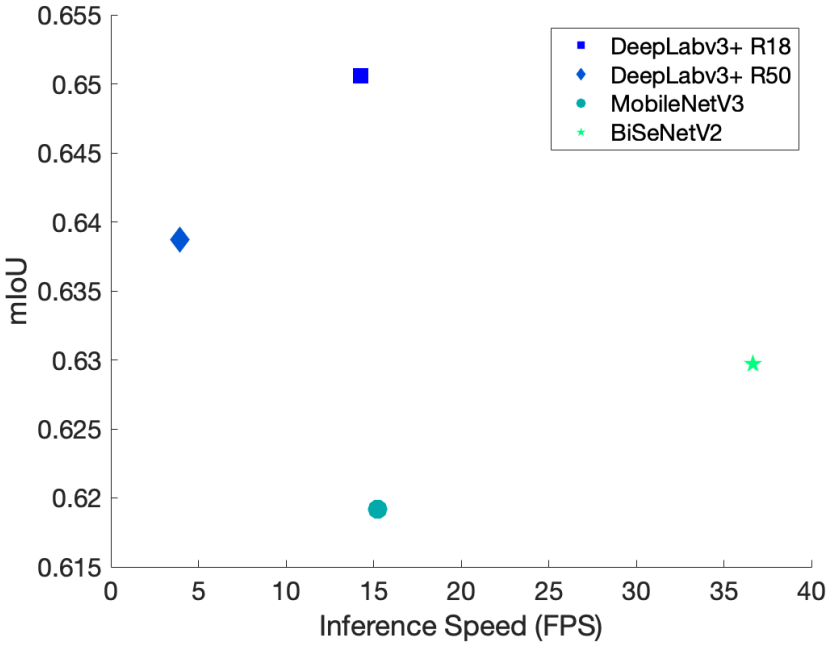


Figure 4.4: The trade-off evaluation between inference speed and segmentation mean Intersection over Union (mIoU) for state-of-the-art segmentation network[52, 19, 7]

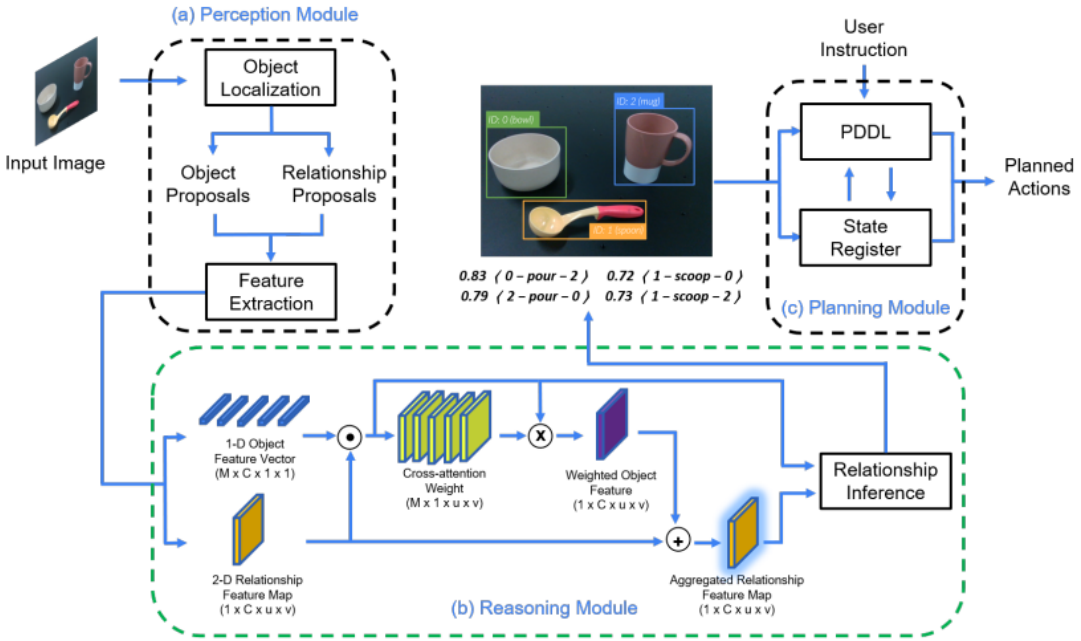


Figure 4.5: The whole pipeline of relationship-oriented semantic scene understanding. Overview: the pipeline first uses (a) perception module to localize all possible object and relationship proposals. A novel relationship-attention the mechanism in the (b) reasoning module is then deployed to aggregate the relationship feature map and infer the relationships along with their probabilities. Finally, the output from the last step seeds the (c) planning module for goal-oriented, multi-step manipulation task planning according to user instructions.

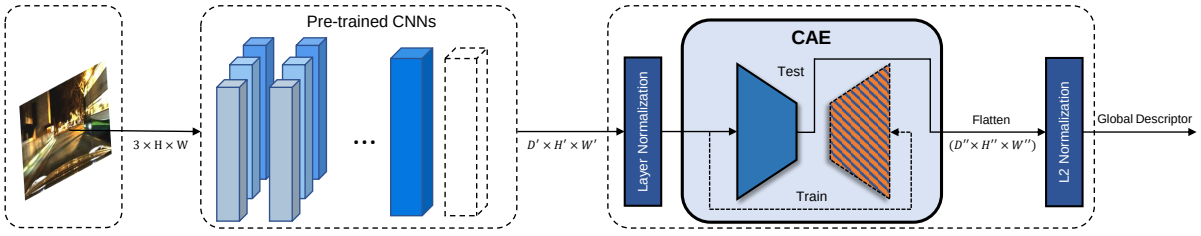


Figure 4.6: The detailed pipeline of [49]. Given an image with $3 \times H \times W$, CNNs extract the local feature map X_i with $D^{'} \times H^{'} \times W^{'}$. The CNNs are classification pre-trained or VPR-trained, e.g. AlexNet, VGG16. Both are cut at the last convolutional layer (conv5), before ReLU. In the training time, CAE is trained unsupervised by a reconstruction loss. In the test time, the decoder part of CAE is not involved and the encoder part is kept to compress the normalized feature map and produce a low-dimensional global descriptor with $D^{''} \times H^{''} \times W^{''}$. The global descriptor is then flattened and L2 normalized.

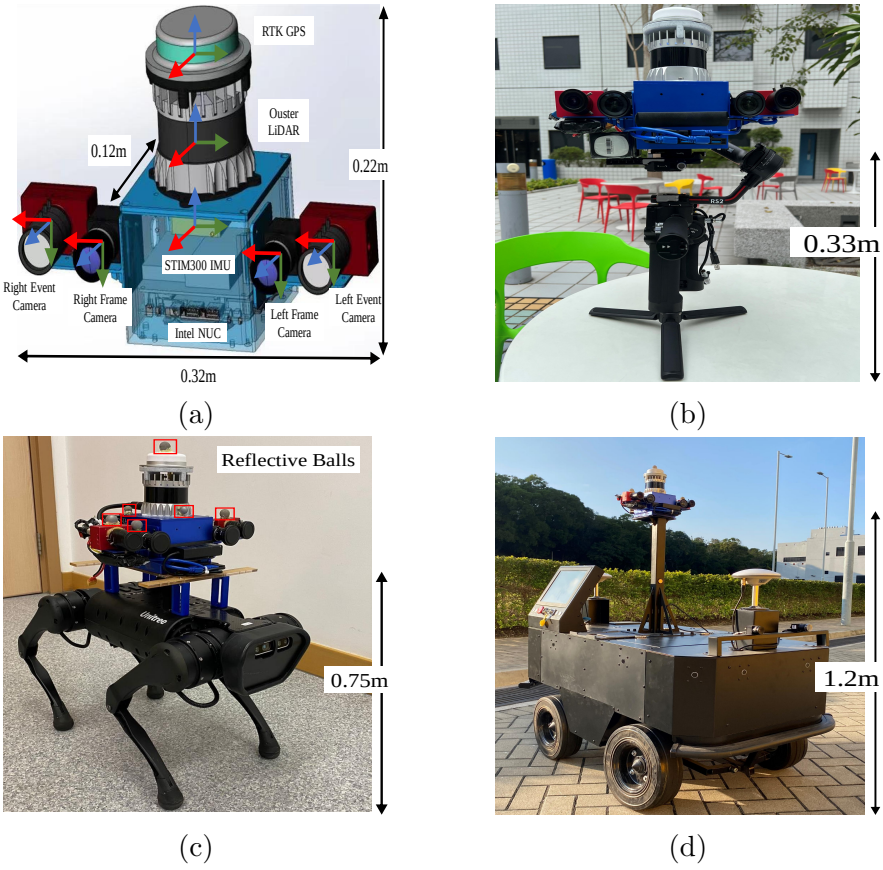


Figure 4.7: The multi-sensor device and data collection platform: (a) CAD model of the sensor rig, where axis directions are colored: red: X, green: Y, blue: Z. The sensor rig is rigidly mounted on (b) a gimbal stabilizer, (c) a quadruped robot, and (d) an apollo autonomous vehicle.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[3] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016.

[4] Yixi Cai, Wei Xu, and Fu Zhang. ikd-tree: An incremental kd tree for robotic applications. *arXiv preprint arXiv:2102.10808*, 2021.

[5] Nicholas Carlevaris-Bianco, Arash K. Ushani, and Ryan M. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *International Journal of Robotics Research*, 35(9):1023–1035, 2015.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[8] Lee Clement, Jonathan Kelly, and Timothy D Barfoot. Robust monocular visual teach and repeat aided by local ground planarity and color-constant imagery. *Journal of field robotics*, 34(1):74–97, 2017.

[9] Dominic Dall'Osto, Tobias Fischer, and Michael Milford. Fast and robust bio-inspired teach and repeat navigation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 500–507. IEEE, 2021.

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[11] Marius Fehr, Fadri Furrer, Ivan Dryanovski, Jürgen Sturm, Igor Gilitschenski, Roland Siegwart, and Cesar Cadena. Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 5237–5244. IEEE, 2017.

[12] Paul Furgale and Timothy D. Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 2010.

[13] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. Sparse-to-dense hypercolumn matching for long-term visual localization. In *2019 International Conference on 3D Vision (3DV)*, pages 513–523. IEEE, 2019.

[14] Mona Gridseth and Timothy D Barfoot. Keeping an eye on things: Deep learned features for long-term visual localization. *IEEE Robotics and Automation Letters*, 7(2):1016–1023, 2021.

[15] Margarita Grinvald, Federico Tombari, Roland Siegwart, and Juan Nieto. Tsdf++: A multi-object formulation for dynamic object tracking and reconstruction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14192–14198. IEEE, 2021.

[16] A. Handa, T. Whelan, J.B. McDonald, and A.J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014.

[17] Mina Henein, Jun Zhang, Robert Mahony, and Viorela Ila. Dynamic slam: The need for speed. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2123–2129. IEEE, 2020.

[18] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous robots*, 34(3):189–206, 2013.

[19] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[20] Huaiyang Huang, Yuxiang Sun, Haoyang Ye, and Ming Liu. Metric monocular localization using signed distance fields. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1195–1201, 2019.

[21] Jianhao Jiao, Hexiang Wei, Tianshuai Hu, Xiangcheng Hu, Yilong Zhu, Zhijian He, Jin Wu, Jingwen Yu, Xupeng Xie, Huaiyang Huang, et al. Fusionportable: A multi-sensor campus-scene dataset for evaluation of localization and mapping accuracy on diverse platforms. *arXiv preprint arXiv:2208.11865*, 2022.

[22] Giseop Kim and Ayoung Kim. Remove, then revert: Static point cloud map construction using multiresolution range images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10758–10765. IEEE, 2020.

[23] Giseop Kim and Ayoung Kim. Lt-mapper: A modular framework for lidar-based lifelong mapping. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7995–8002, 2022.

[24] Tomáš Krajník, Filip Majer, Lucie Halodová, and Tomáš Vintr. Navigation without localisation: reliable teach and repeat based on the convergence theorem. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1657–1664. IEEE, 2018.

[25] Mans Larsson, Erik Stenborg, Carl Toft, Lars Hammarstrand, Torsten Sattler, and Fredrik Kahl. Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 31–41, 2019.

[26] Shigang Li and Akira Hayashi. Robot navigation in outdoor environments by using gps information and panoramic views. In *Proceedings. 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems. Innovations in Theory, Practice and Applications (Cat. No. 98CH36190)*, volume 1, pages 570–575. IEEE, 1998.

[27] Tianyu Liu, Qinghai Liao, Lu Gan, Fulong Ma, Jie Cheng, Xupeng Xie, Zhe Wang, Yingbing Chen, Yilong Zhu, Shuyang Zhang, et al. Hercules: An autonomous logistic vehicle for contact-less goods transportation during the covid-19 pandemic.

[28] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017.

[29] Colin McManus, Paul Furgale, Braden Stenning, and Timothy D Barfoot. Visual teach and repeat using appearance-based lidar. In *2012 IEEE International Conference on Robotics and Automation*, pages 389–396. IEEE, 2012.

[30] Tayyab Naseer, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard. Semantics-aware visual localization under challenging perceptual conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2614–2620, 2017.

[31] Matías Nitsche, Facundo Pessacg, and Javier Civera. Visual-inertial teach and repeat. *Robotics and Autonomous Systems*, 131:103577, 2020.

[32] Helen Oleynikova, Zachary Taylor, Marius Fehr, Roland Siegwart, and Juan Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1366–1373. IEEE, 2017.

[33] Michael Paton, Kirk MacTavish, Michael Warren, and Timothy D Barfoot. Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1918–1925. IEEE, 2016.

[34] Jingxing Qian, Veronica Chatrath, Jun Yang, James Servos, Angela P Schoellig, and Steven L Waslander. Pocd: Probabilistic object-level change detection and volumetric mapping in semi-static scenes. *arXiv preprint arXiv:2205.01202*, 2022.

[35] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019.

[36] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020.

[37] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[38] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019.

[39] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

[40] Lukas Schmid, Jeffrey Delmerico, Johannes L Schönberger, Juan Nieto, Marc Pollefeys, Roland Siegwart, and Cesar Cadena. Panoptic multi-tsdfs: a flexible representation for online multi-resolution volumetric mapping and long-term dynamic scene consistency. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8018–8024. IEEE, 2022.

[41] Christoph Sprunk, Gian Diego Tipaldi, Andrea Cherubini, and Wolfram Burgard. Lidar-based teach-and-repeat of mobile robot trajectories. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3144–3149. IEEE, 2013.

[42] Erik Stenborg, Carl Toft, and Lars Hammarstrand. Long-term visual localization using semantically segmented images. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6484–6490. IEEE, 2018.

[43] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[44] Li Sun, Marwan Taher, Christopher Wild, Cheng Zhao, Yu Zhang, Filip Majer, Zhi Yan, Tomáš Krajník, Tony Prescott, and Tom Duckett. Robust and long-term monocular teach and repeat navigation using a single-experience map. In *2021*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2635–2642. IEEE, 2021.

[45] Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, and Akihiko Torii. Is this the right place? geometric-semantic pose verification for indoor visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4373–4383, 2019.

[46] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual place recognition with repetitive structures. 2015.

[47] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019.

[48] Linhui Xiao, Jinge Wang, Xiaosong Qiu, Zheng Rong, and Xudong Zou. Dynamic-slam: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robotics and Autonomous Systems*, 117:1–16, 2019.

[49] Hanjing Ye, Weinan Chen, Jingwen Yu, Li He, Yisheng Guan, and Hong Zhang. Condition-invariant and compact visual place description by convolutional autoencoder. *arXiv preprint arXiv:2204.07350*, 2022.

[50] Haoyang Ye, Huaiyang Huang, and Ming Liu. Monocular direct sparse localization in a prior 3d surfel map. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8892–8898, 2020.

[51] Peng Yin, Abulikemu Abuduweili, Shiqi Zhao, Changliu Liu, and Sebastian Scherer. Bioslam: A bio-inspired lifelong memory system for general place recognition. *IEEE Transactions on Robotics, Conditional Accepted*, 2022.

[52] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021.

[53] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments. In *2018 IEEE/RSJ*

International Conference on Intelligent Robots and Systems (IROS), pages 1168–1174, 2018.

[54] Yang Yu, Peng Yun, Bohuan Xue, Jianhao Jiao, Rui Fan, and Ming Liu. Accurate and robust visual localization system in large-scale appearance-changing environments. *IEEE/ASME Transactions on Mechatronics*, 2022.

[55] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, volume 2, pages 1–9. Berkeley, CA, 2014.