2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

A Hierarchical Framework for Collaborative Probabilistic Semantic Mapping

Yufeng Yue*, Chunyang Zhao, Ruilin Li, Chule Yang, Jun Zhang, Mingxing Wen
Yuanzhe Wang and Danwei Wang

Abstract— **Performing collaborative semantic mapping is a critical challenge for cooperative robots to maintain a comprehensive contextual understanding of the surroundings. Most of the existing work either focus on single robot semantic mapping or collaborative geometry mapping. In this paper, a novel hierarchical collaborative probabilistic semantic mapping framework is proposed, where the problem is formulated in a distributed setting. The key novelty of this work is the mathematical modeling of the overall collaborative semantic mapping problem and the derivation of its probability decomposition. In the single robot level, the semantic point cloud is obtained based on heterogeneous sensor fusion model and is used to generate local semantic maps. Since the voxel correspondence is unknown in collaborative robots level, an Expectation-Maximization approach is proposed to estimate the hidden data association, where Bayesian rule is applied to perform semantic and occupancy probability update. The experimental results show the high quality global semantic map, demonstrating the accuracy and utility of 3D semantic map fusion algorithm in real missions.**

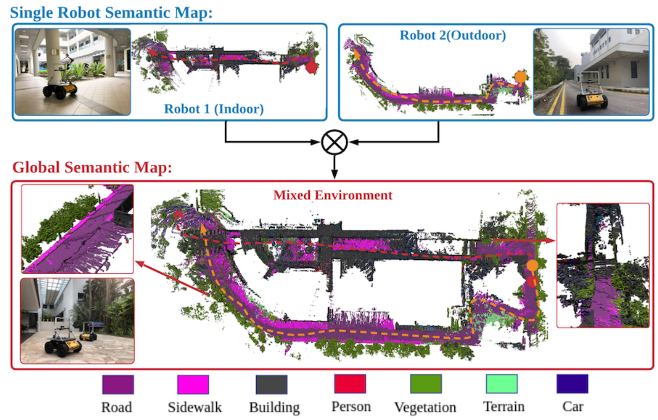

Fig. 1. A demonstration of collaborative semantic mapping in mixed indoor-outdoor environment. Red trajectory is robot 1, and orange trajectory is robot 2. Top: local semantic maps generated by two robots. Bottom: collaboratively generated global semantic map.

I. INTRODUCTION

Autonomous robots need to include semantic information to operate reliably and efficiently. Current 3D geometry mapping approaches only contain geometry information and limit the robot's application in high-level tasks. Semantic mapping consists of geometry mapping [1] and semantic information [2], which is crucial for path planning and navigation [3]. Recently, single robot semantic mapping has attracted many attention [4], [5]. For collaborative robots, approaches have been proposed to perform collaborative geometry mapping [6], [7]. This paper moves one step forward by focusing on collaborative probabilistic semantic mapping.

Collaborative 3D semantic mapping attempts to fuse the semantic maps generated by individual robots (i.e., local maps) into a global semantic map [8]. Due to limited sensing capabilities, each robot only has partial information of the surroundings. Therefore, it is crucial to share perceptual semantic information among all robots to gain a comprehensive understanding of the environment. Recently, studies on map fusion have been addressed by different areas. Map stitching [9] and sub-map assembling [10] methods are widely applied in single robot mapping. However, they are essentially methods for single robot. This paper differs from them by considering unknown data association between voxels in local maps generated by different robots. Single robot mapping does not consider communication, but distributed setup is a key factor in ensuring system utility under narrow bandwidth and limited computing power. Compared to collaborative geometry mapping [6], semantic information update and integration bring another layer of challenge. In short, comprehensive analysis, modeling and implementation of collaborative semantic mapping methods have not been studied in-depth.

This paper proposes a novel collaborative semantic mapping framework. The key novelty of this work is the mathematical modeling of the overall collaborative semantic mapping problem and the derivation of its probability decomposition, as well as the experimental implementation in real environment. The main contributions are listed as follows:

- A hierarchical collaborative probabilistic semantic mapping framework is proposed to address the problems in both single-robot and collaborative robots levels.
- An Expectation-Maximization (EM) approach is proposed to estimate the hidden data association between voxels in local maps, where Bayesian rule is applied to perform semantic and occupancy probability update.
- Collaborative robots system is developed to perform scene understanding and global semantic mapping, demonstrating the accuracy and utility in real missions.

The rest of the paper is organized as follows. Section II reviews the related work. Section III gives an overview of the

The research was partially supported by the ST Engineering-NTU Corporate Lab through the NRF corporate lab@university scheme. (Corresponding Author: Yufeng Yue)

Yufeng Yue*, Chunyang Zhao, Ruilin Li, Jun Zhang, Mingxing Wen, Yuanzhe Wang, Danwei Wang are with School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore (email: yyue001@e.ntu.edu.sg)

Chule Yang is with National Innovation Institute of Defense Technology, Military Academy of Sciences, Beijing, 100071, China

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

proposed framework. Section IV explains the theoretical basis for collaborative semantic mapping. Section V shows the experimental procedures and results. Section VI concludes the paper with a discussion on future work.

II. Related Work

This section first gives an overview of semantic segmentation and single robot semantic mapping, and then introduces recent work on collaborative geometry mapping.

A. Single Robot Semantic Mapping

Simultaneous Localization and Mapping (SLAM) is initially proposed to address robot localization and mapping problem in unknown environment [11]. As the availability of high-quality 3D sensors increases, a framework of generating volumetric 3D environment model is proposed in [1]. However, traditional SLAM only restores the geometry information of the surroundings. With the fast development of deep learning, it becomes possible to obtain the semantic information by applying CNN models. Deeplab [12], at present, is one of the best model that has excellent performance in semantic segmentation.

Recently, algorithms have been proposed to address single robot semantic mapping problem. An incremental semantic 3D mapping system for large-scale environments using a scrolling occupancy grid is proposed in [13]. To improve the efficiency, octree based multi-label 3D semantic mapping algorithm is proposed in [5]. Since semantic information is also important for localization, [14] integrates the semantic information into SLAM approach. More recently, a semantic SLAM system [15] is presented that uses object-level entities to construct semantic objects and integrate them into the semantic SLAM framework. To perform mapping in dynamic environments [16], a stereo-based dense mapping algorithm is proposed [17]. Regarding the reconstruction of moving objects, [18] incrementally fuses sensor observations into a consistent semantic map. Aforementioned approaches promote the development of single robot semantic mapping, but collaborative semantic mapping is still an open problem.

B. Collaborative Geometry Map Fusion

Various efficient solutions have been developed to solve collaborative geometry mapping problem. In [19], different solutions for collaborative mapping are summarized. These techniques range from front-end 2D topology extraction [20], sparse feature descriptors [21], full 3D dense map registration [22], to back-end optimization of splitting maps into sub-maps [23] and spatial constraints graph structures [24].

Global mapping in real environment requires careful selection of map types based on actual conditions to balance the requirements of detailed 3D mapping, transmission through limited communication, and practicality for robust map fusion [20]. Data fusion between multiple robots can be grouped into three different types, raw sensor data [24], volumetric maps [22] and topological maps [20]. Metric grid maps such as 3D volumetric maps [1] model the environment into probabilistic form with sufficient details. In this paper, we adopt the efficient 3D octree representation proposed by [1]. To sum up, the above algorithms contribute a lot to collaborative mapping problem. However, they did not consider the issue of semantic map fusion. Therefore, a new system framework and theoretical formula are needed.

III. System Framework

A. The Framework of Hierarchical Semantic 3D Mapping

The overview of the system architecture is depicted in Figure 2, where the framework consists of three modules: *multimodal semantic information fusion, single robot semantic mapping and collaborative semantic map fusion*. Multimodal semantic information fusion is at the perception level, the robots generate semantic point clouds based on sensor calibration and image segmentation. Single robot semantic mapping is at the local mapping level, each robot builds its own semantic map locally. Collaborative semantic mapping fusion is at the global map level, the robots communicate with each other to transmit and fuse the local semantic maps into a global semantic map.

B. Centralized Problem Formulation

Consider a group of robots moving through an unknown environment and attempt to map out the surroundings. The problem can be defined as follows:

Centralized Definition: *Given a group of robots* $\mathcal{R} \triangleq \{r\}_{1:R}$, *the objective is to estimate the global semantic map* M_t *given camera observations* $I_{1:t}^{(\mathcal{R})}$, *3D laser observations* $L_{1:t}^{(\mathcal{R})}$ *and robot trajectory* $x_{1:t}^{(\mathcal{R})}$.

$$p(M_t|I_{1:t}^{(\mathcal{R})}, L_{1:t}^{(\mathcal{R})}, x_{1:t}^{(\mathcal{R})}) \tag{1}$$

The global semantic map $M = \{M_i\}_1^N$ consists of a set of voxels. Each voxel $M_i = (M_x^i, M_y^i, M_z^i, v_i, o_i)$ is defined as a tuple that includes the position of the extracted voxel center $m_i = (M_x^i, M_y^i, M_z^i)$, the occupancy probability value v_i, the semantic label o_i. The label o_i is assumed to come from a finite set of discrete class labels: $S = \{1, 2, \cdots, k, \cdots\}$. We have $x_t \in SE(3)$ in 3D, $I_t \in \mathbb{R}^2$ in 2D and $L_t \in \mathbb{R}^3$ in 3D.

In the centralized setting, the problem corresponds to calculate the Maximum A Posterior (MAP) estimation problem (1). Assuming the communication is perfect, all the robots can share their latest sensor observations $L_{1:t}^{(\mathcal{R})}$ and $I_{1:t}^{(\mathcal{R})}$ to the control station in real time. However, it is challenging to transfer large size raw sensor data under limited bandwidth.

C. Distributed Definition

For distributed framework in constraint environment, each robot r only has access to its own raw sensor observations. In this case, the paper introduces the single robot layer to perform local semantic mapping, which is an intermediate level that acts as perception and commutation node. In order to improve network efficiency and robustness, only the generated local semantic maps are allocated in the group of

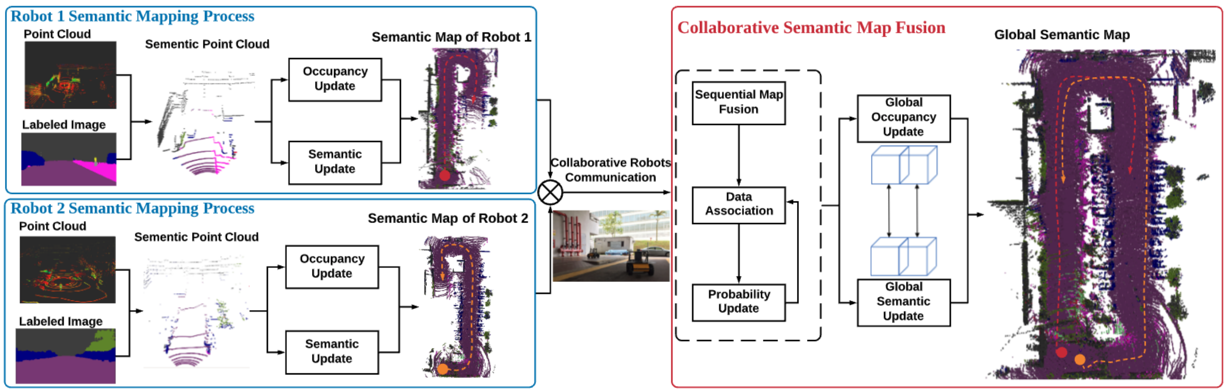

Fig. 2. The framework of hierarchical collaborative semantic 3D mapping. Red trajectory is robot 1, and orange trajectory is robot 2.

robots. The collaborative robots level turns into estimating the global semantic map given the outputs from single robot.

Single Robot Level Definition: *For each robot r, the objective is to estimate its local semantic map $m_t^{(r)}$ given its camera observations $I_{1:t}^{(r)}$, 3D laser observations $L_{1:t}^{(r)}$ and robot trajectory $x_{1:t}^{(r)}$.*

$$p(m_t^{(r)} | I_{1:t}^{(r)}, L_{1:t}^{(r)}, x_{1:t}^{(r)}) \tag{2}$$

In single robot level, raw sensor data $I_{1:t}^{(r)}$ and $L_{1:t}^{(r)}$ serves as the input. First, the multimodal information fusion algorithm is proposed to generate semantic point cloud (see IV-A). Then, the semantic point cloud is used to generate local semantic map based on Bayes rule, and the output is single robot semantic map $m_t^{(r)}$ (see IV-B).

Collaborative Robots Level Definition: *The objective of distributed collaborative semantic mapping is to estimate global semantic map M_r under a fully distributed network, given local maps $m_t^{(r,\mathcal{R}_r)}$ from neighboring robots $\mathcal{R}_r$.*

$$p(M_r | m_t^{(r)}, \Phi_{r_n \in \mathcal{R}_r}(m_t^{(r_n)})) \tag{3}$$

In collaborative robots level, each robot r is assumed to have estimated its semantic map $m_t^{(r)}$ based on local observations. Robot r receives the local maps $m_t^{(r_n)}$ from all the nearby robots $r_n \in \mathcal{R}_r$, where $\mathcal{R} = \{r \cup \mathcal{R}_r\}$ if the communication covers all robots. Given the received local maps, Expectation-Maximization algorithm is applied to jointly estimate the hidden data association and update the probability distribution (see IV-C).

IV. Collaborative Semantic 3D Mapping

This section presents detailed distributed collaborative semantic mapping framework, which is divided into three subsections, i.e., multimodal semantic information fusion, single robot semantic mapping and collaborative semantic map fusion.

A. Multimodal Semantic Information Fusion

Our semantic 3D mapping method uses a 3D LiDAR and a camera as its main sensors. The images are processed by

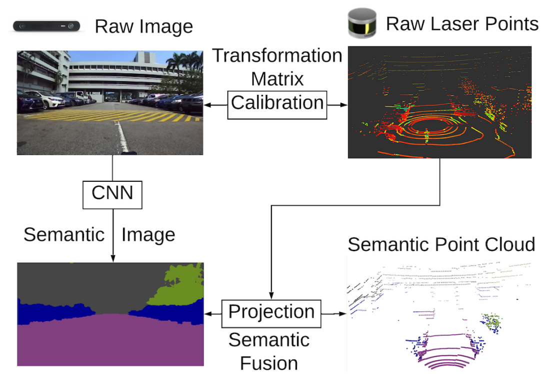

Fig. 3. Flowchart of multimodal semantic information fusion.

Deeplab model [25] to obtain semantic classes. Here, the Deeplab model outputs 19 semantic classes. After image segmentation, the raw image I is transferred to labeled semantic image I_s.

Applying the algorithm we proposed in [26], we can obtain the projection matrix P_L^I, with which the 3D point cloud can be projected to 2D image plane. Meanwhile, using the inverse projection matrix $P_L^{I^{-1}}$, the semantic pixel point $I_s(u,v)$ can be re-projected to 3D point cloud $L(x,y,z)$. Thus, the 3D point $L(x,y,z)$ will have semantic label $L_s(x,y,z)$ (see (4)).

$$L_s(x,y,z) = P_L^{I^{-1}} * I_s(u,v) \tag{4}$$

In the semantic mapping process, 3D geometry coordinate $L_{1:t}^{(r)}$ and the corresponding 19D semantic labels $L_{s_{1:t}}^{(r)}$ will serve as the input. Therefore, (2) can be rewritten as:

$$p(m_t^{(r)} | L_{s_{1:t}}^{(r)}, L_{1:t}^{(r)}, x_{1:t}^{(r)}) \tag{5}$$

As each laser scan is with respect to its local sensor frame. Each point cloud is transformed using the odometry $x_{1:t}$ for 3D map generation, where $\hat{L}_t = x_t * L_t$, $\hat{L}_{st} = x_t * L_{st}$, then we have $p(m_t^{(r)} | L_{s_{1:t}}^{(r)}, L_{1:t}^{(r)}, x_{1:t}^{(r)}) = p(m_t^{(r)} | \hat{L}_{1:t}^{(r)}, \hat{L}_{s_{1:t}}^{(r)})$.

B. Single Robot Semantic Mapping

1) Voxel Occupancy and Label Update: Here, the denotation of $m_t^{(r)}$ is simplified as m_r. For the partial map, we define $m_r = \{m_r^i\}_{i=1}^{N_{m_r}}$, $m_{r_n} = \{m_{r_n}^j\}_{j=1}^{N_{m_{r_n}}}$, where N_{m_r} and $N_{m_{r_n}}$ are the number of voxels of partial maps. The probability of a leaf node m_r^i is updated by applying the Bayes rule. Since the voxel-wise correspondences are usually assumed to be independent, the probability density function of $p(m_t^{(r)}|I_{1:t}^{(r)}, L_{1:t}^{(r)}, x_{1:t}^{(r)})$ can be factorized as:

$$p(m_r|\hat{L}_{s\,1:t}^{(r)}, \hat{L}_{1:t}^{(r)}) = \prod_{i=1}^{n} p(m_r^i|\hat{L}_{s\,1:t}^{(r)}, \hat{L}_{1:t}^{(r)}) \tag{6}$$

For semantic mapping, we are updating occupancy probability and semantic label probability simultaneously. Based on conditional independence, each voxel is updated separately with occupancy update model $p(v_r^i|\hat{L}_{1:t}^{(r)})$ and semantic update model $p(o_r^i|\hat{L}_{s\,1:t}^{(r)})$, then (6) is rewritten as (7).

$$\prod_{i=1}^{n} p(v_r^i, o_r^i|\hat{L}_{s\,1:t}^{(r)}, \hat{L}_{1:t}^{(r)}) = \prod_{i=1}^{n} \underbrace{p(v_r^i|\hat{L}_{1:t}^{(r)})}_{\text{Occupancy Update}} \cdot \underbrace{p(o_r^i|\hat{L}_{s\,1:t}^{(r)})}_{\text{Semantic Update}} \tag{7}$$

The occupancy probability is recursively updated given the incoming 3D point cloud $\hat{L}_{1:t}^{(r)}$, where $p(v_r^i)$ denotes the initial occupancy probability of each voxel and is set as 0.5. Then, occupancy probability is updated in (8).

$$p(v_r^i|\hat{L}_{1:t}^{(r)}) = [1 + \frac{1 - p(v_r^i|\hat{L}_t^{(r)})}{p(v_r^i|\hat{L}_t^{(r)})} \frac{1 - p(v_r^i|\hat{L}_{1:t-1}^{(r)})}{p(v_r^i|\hat{L}_{1:t-1}^{(r)})} \frac{p(v_r^i)}{1 - p(v_r^i)}]^{-1} \tag{8}$$

The semantic probability is recursively updated given the incoming set of 19 semantic labels $\hat{L}_{s\,1:t}^{(r)} = \{\hat{L}_{s\,1:t}^{(r)}(k)\}_{k=1}^{19}$. Then, $p(o_r^i|\hat{L}_{s\,1:t}^{(r)})$ denotes the probability of updated label class (k=1:19). $p(o_r^i)$ denotes the initial semantic probability of each voxel and is set as $\frac{1}{19}$. Then, semantic probability of each class is updated in (9).

$$p(o_r^i|\hat{L}_{s\,1:t}^{(r)}) = [1 + \frac{1 - p(o_r^i|\hat{L}_{s\,t}^{(r)})}{p(o_r^i|\hat{L}_{s\,t}^{(r)})} \frac{1 - p(o_r^i|\hat{L}_{s\,1:t-1}^{(r)})}{p(o_r^i|\hat{L}_{s\,1:t-1}^{(r)})} \frac{p(o_r^i)}{1 - p(o_r^i)}]^{-1} \tag{9}$$

After fusion, the class that corresponds to maximum probability is assign as the label of the voxel. The process of single robot semantic map update is shown in Fig. 4(a-b).

C. Collaborative Semantic Map Fusion

1) Sequential Semantic Map Fusion: Under limited bandwidth condition, maps are generated and transmitted sequentially in a certain time interval. This results in robot r receiving its neighboring robots' maps in some permutation π. Hence, the map fusion is performed serially, where a certain threshold is satisfied to trigger a pair-wise map sharing and fusion. Then, Eq. (3) can be factorized into Eq. (10). Initially, the relative transformation matrix T_{r,r_n} between m_r and m_{r_n} is unknown. In this paper, the map matching algorithm we have proposed in [27] is applied to estimate T_{r,r_n}. Then, the neighboring robot map m_{r_n} is

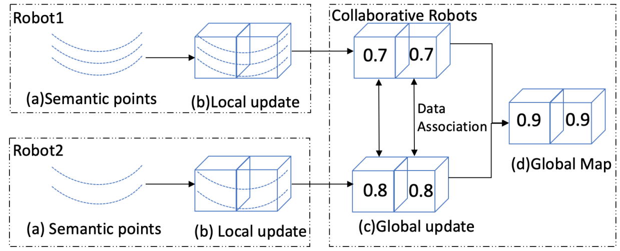

Fig. 4. An example of local and global semantic map update process. The number here represents semantic probability, indicating the semantic update process.

transformed to coordinate frame of robot r by transformation function $\Phi(T_{r,r_n}, m_{r_n})$, which is simplified as $\Phi(m_{r_n})$.

$$\begin{aligned} &p(M_r|m_r, \Phi_{r_n \in \mathcal{R}_r}(m_{r_n})) \\ &\qquad = \prod_{r_n=1}^{N_r} p(M_r^{\pi(1:r_n)}|M_{r_n}^{\pi(1:r_n-1)}, \Phi(m_{\pi(r_n)})) \end{aligned} \tag{10}$$

Note that $M_r^{\pi(1,r_n)}$ is the partial global map of robot r generated by fusing the maps of robots $\pi(1)$ to $\pi(r_n) \in \mathcal{R}_r$ with m_r after the reception of map $m_{\pi(r_n)}$ from neighboring robot $\pi(r_n) \in M_r^{\mathcal{R}_r}$. The final global fused map $M_r^{\mathcal{R}_r}$ of robot r, can be retrieved after the end of this serial process, hence $M_r^{\mathcal{R}_r} = M_r^{\pi(1,Nr)}$.

2) Global Map Occupancy and Label Update: The key difference between local and global semantic map update comes at two hands. First, the input of single robot mapping is raw sensor observation (Fig. 4(a)), while the input of global semantic mapping is local semantic maps (Fig. 4(c)). Since voxel correspondence is unknown, and we need to establish the data association relationship before fusion. Second, the same object can be observed in different perspectives by different robots, the voxels representing the same object can have different semantic class. Hence it is vital to come up with a strategy to consider the dissimilarities and fuse them into the global semantic map (Fig. 4(d)).

For simplification, we denote the latest incoming neighboring robot map as m_{r_n} and current robot map as m_r. To estimate the hidden data association between voxels, Expectation Maximization (EM) is applied, which is an algorithm well-suited for models containing latent variables. There are two steps for EM: the E-step efficiently estimates the hidden variables by evaluating the expectation, while the M-step updates the global map probability given the corresponding voxel pairs. A binary variable d is introduced to represent the corresponding voxels, where $d_{i,j} = 1$ if m_r^i corresponds to $m_{r_n}^j$ and equals 0 otherwise.

$$p(M_r|m_r, m_{r_n}) = \prod_{i=1} p(M_r^k|m_r^i, m_{r_n}^j) \tag{11}$$

$$= \prod_{i=1} \sum_{j}^{n} p(M_r^k, d_{i,j}|m_r^i, m_{r_n}^j) \tag{12}$$

$$= \prod_{i} \sum_{j} \underbrace{p(d_{i,j}|m_r^i, m_{r_n}^j)}_{\text{E step:data association}} \cdot \underbrace{p(M_r^k|d_{i,j}, m_r^i, m_{r_n}^j)}_{\text{M step:probability update}} \tag{13}$$

E-step establishes the correspondence by calculating the minimum relative distance. Here, we define a 5D descriptor as $\{m_i^x, m_i^y, m_i^z, v_{m_i}, o_{m_i}\}$, which includes the center coordinate of voxel ($3D$), occupancy probability ($1D$) and semantic labeling ($1D$). The corresponding voxel is calculated by finding the nearest neighborhood as formulated in (14).

$$\begin{aligned} p(d_{i,j}|m_r^i, m_{r_n}^j) &= d_e(m_r^i, m_{r_n}^j) + d_v(v_{m_r^i}, v_{m_{r_n}^j}) + o_v(o_{m_r^i}, o_{m_{r_n}^j}) \\ &= ||m_r^i - m_{r_n}^j||^2 + ||v_{m_r^i} - v_{m_s^j}||^2 + ||o_{m_r^i} - o_{m_s^j}||^2 \end{aligned} \tag{14}$$

Given the data association $d_{i,j}$, the next step is to update corresponding voxels in collaborative robots level. The occupancy probability and semantic label probability are updated separately.

$$p(M_r^k|d_{i,j}, m_r^i, m_{r_n}^j) \tag{15}$$

$$= \prod_{i=1}^{n} p(v(M_r^k), o(M_r^k)|d_{i,j}, v(m_r^i), v(m_{r_n}^j), o(m_r^i), o(m_{r_n}^j)) \tag{16}$$

$$= \prod_{i=1}^{n} \underbrace{p(v(M_r^k)|d_{i,j}, v(m_r^i), v(m_{r_n}^j))}_{\text{Global Occupancy Update}} \cdot \underbrace{p(o(M_r^k)|d_{i,j}, o(m_r^i), o(m_{r_n}^j))}_{\text{Global Semantic Update}} \tag{17}$$

In single robot occupancy updating (8), the input is a set of raw semantic points. However, the input in collaborative robots level is the the probability value of local semantic map. Each voxel can be regraded as a Gaussian distribution, which is located at the center of the voxel and identical in each direction. In this case, the fusion at collaborative robots level is the integration of two Gaussian distributions. Then, the global occupancy update is formulated in (18).

$$\begin{aligned} &p(v(M_r^k)|d_{i,j}, v(m_r^i), v(m_{r_n}^j)) \\ &= [1 + \frac{1 - p(v(m_{r_n}^j))}{p(v(m_{r_n}^j))} \frac{1 - p(v(m_r^i))}{p(v(m_r^i))} \frac{p(v(M_r^k))}{1 - p(M_r^k)}]^{-1} \end{aligned} \tag{18}$$

For semantic probability update, the input is 19 class probabilities of each voxel in single robot. The initial semantic probability of each global map voxel $p(o(M_r^k))$ is $\frac{1}{19}$.

$$\begin{aligned} &p(o(M_r^k)|d_{i,j}, o(m_r^i), o(m_{r_n}^j)) \\ &= [1 + \frac{1 - p(o(m_{r_n}^j))}{p(o(m_{r_n}^j))} \frac{1 - p(o(m_r^i))}{p(o(m_r^i))} \frac{p(o(M_r^k))}{1 - p(o(M_r^k))}]^{-1} \end{aligned} \tag{19}$$

The semantic probability of the most likely class $p(o((M_r^k)), max)$ for each node M_r^k is computed as follows:

$$p(o(M_r^k)), max) = \arg\max[p(o(M_r^k), 1), \ldots, p(o(M_r^k), 19)] \tag{20}$$

V. Experimental Results

Experiments are conducted by operating two robots in two different scenarios. All algorithms are executed on the ROS platform. Two robot platforms (Husky Clearpath) are equipped with the Intel Core i7-6700HQ CPU @ 2.60GHz CPU and the Nvidia GeForce RTX 2060 @ 6GB RAM GPU. Each robot is equipped with a 3D Velodyne LiDAR and visual camera. The semantic segmentation model is trained by using the cityscapes dataset [28]. LOAM [29] is applied for single robot pose estimation, HPFF [27] is implemented to estimate relative transformation between robots, and Octomap [1] is utilized for basic 3D geometry mapping at the resolution of 0.2m. The communication between robots is established by long range Wi-Fi with limited bandwidth.

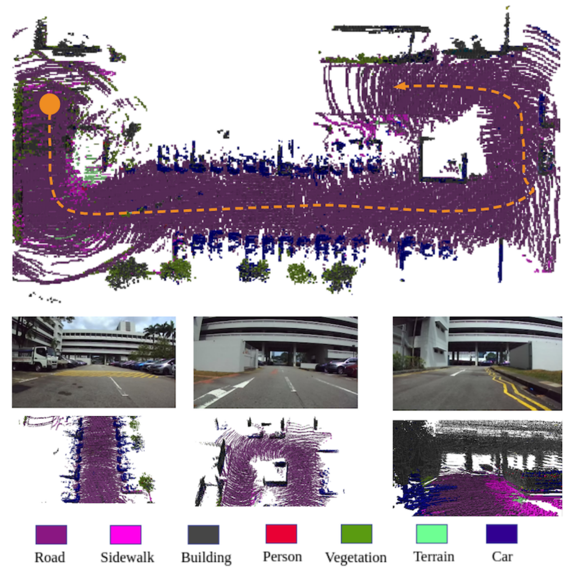

Fig. 5. Robot 2 semantic map in open carpark. Only the seven classes that appear in the scene are marked.

The experiment environment settings are:

- **Open Carpark**: two Husky robots equipped with Velodyne 3D LiDAR and visual camera in an open carpark.
- **Mixed Environment**: two Husky robots equipped with Velodyne 3D LiDAR and visual camera in an indoor-outdoor mixed environment.

Comparison Baseline: Most of the existing work either focus on single robot semantic mapping or collaborative geometry mapping. As no available work has addressed this problem, we first demonstrate extensive qualitative results of the proposed algorithm. Then, we present quantitative results on the detailed process of how the global map is updated and how the local maps are fused.

A. *Single Robot Semantic Mapping*

To evaluate the multimodal semantic fusion algorithm for single robot semantic mapping, the mapping result is presented in Fig. 5. The figure shows top view of the whole map and three close-up views for each time step, where each image corresponds to the related local semantic map. As shown in Fig. 5, our method successfully integrates and updates the semantic information into 3D semantic map.

B. *Collaborative Semantic Mapping*

The figure depicted in Fig. 6 is at the open carpark in the university. Two robots started their mapping from a common

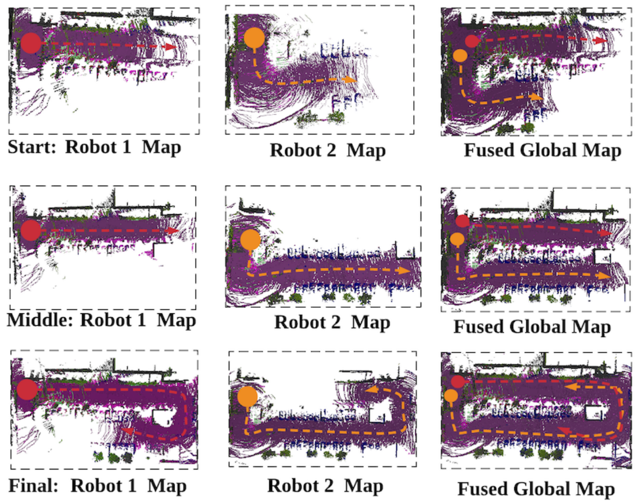

Fig. 6. Collaborative semantic mapping at the start, middle and final of the task in the open carpark.

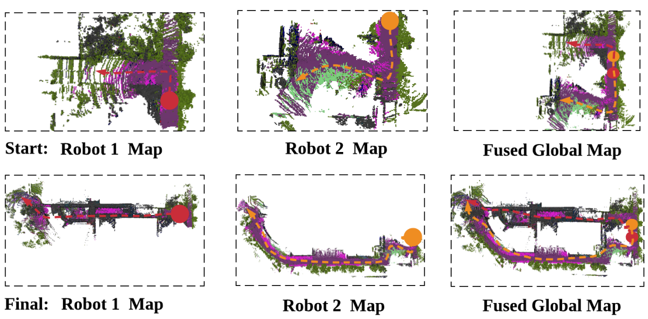

Fig. 7. Collaborative semantic mapping at the start and final of the task in the mixed environment.

neighborhood and explored the environment by traversing two different paths. Fig. 6 shows the fusion of maps at start, middle and final of the mission. For collaborative robots, the algorithm achieves the fusion of maps from two robots. The algorithm combines the map information from two maps to generate an enhanced and more consistent global map.

The mixed environment depicted in Fig. 7 is at the ground floor of a building in the university, where the indoor environment is the lobby, and the outdoor environment is the outside road. The two robots first started nearby in the initial position, it can be seen that the two maps are successfully merged into a global map. After crossing the underpass, two robots encountered at the end position. The overlapped area between the two maps combines the advantages of the two sensors, which produces a more detailed and abundant fused map.

C. Quantitative analysis

Table I demonstrates the fusion of process of a pair of corresponding voxels from two robots. The corresponding pair is estimated by the EM algorithm proposed in IV-C. It is worth noting that the voxel (true label is Car) from robot **1** is assigned with the class Road with a low probability, which is a false negative. The corresponding voxel in the robot **2** has a high confidence level of 0.992 for the true class Car. The voxel label in the global map is assigned to Car after fusion. This shows the superiority of the proposed algorithm for correcting false negative label after fusion. Then, Fig. 8 shows more detailed semantic probability update process of this voxel in global map, specifying the label and its corresponding probability. We can observe the trend of probability increasing of the label Car (blue) and the decreasing in the remaining labels.

TABLE I

GLOBAL MAP LABEL FUSION (CORRECT FALSE NEGATIVE (FN) LABEL)

	Robot1 (FN)		Robot2 (TP)		Global (TP)	
No	Label	Value	Label	Value	Label	Value
1	Road	0.638	Car	0.973	Car	**0.894**
2	Road	0.711	Car	0.985	Car	**0.969**
3	Road	0.834	Car	0.992	Car	**0.977**

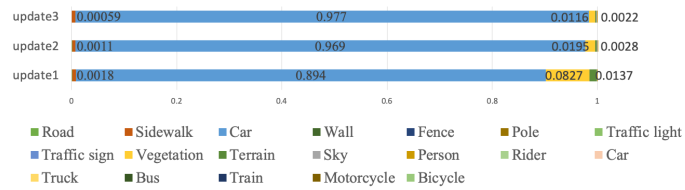

Fig. 8. Semantic probability update process of a voxel in global map.

TABLE II

GLOBAL MAP LABEL FUSION (ENHANCE TRUE POSITIVE (TP) LABEL)

	Robot1 (TP)		Robot2 (TP)		Global (TP)	
No	Label	Value	Label	Value	Label	Value
1	Road	0.695	Road	0.773	Road	**0.784**
2	Road	0.810	Road	0.855	Road	**0.886**
3	Road	0.906	Road	0.924	Road	**0.943**

Table II shows the example of fusion process can enhance semantic probability of true positive label. The corresponding voxels from two local maps indicate the true class label of Road. The fusion integrates the probability information and increases the semantic probability in the global map.

VI. CONCLUSION

This paper has established a hierarchical collaborative probabilistic semantic mapping framework. In the single robot level, semantic point cloud is obtained based on heterogeneous sensor fusion model and used to generate local semantic maps. To achieve collaborative semantic mapping, this paper has provided theoretical basis and algorithm implementation for the global 3D semantic mapping. The results have shown the algorithm is able to establish the data association between voxels. More importantly, the fusion process is able to correct the false label and enhance true label. The overall experimental results have presented high quality global semantic maps, demonstrated the accuracy and utility of the framework. In the future, collaborative robots localization and place recognition can be performed based on the semantic maps, which shall lead to a significant improvement.

REFERENCES

[1] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees," *Autonomous Robots*, 2013.

[2] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[3] Y. Wang, M. Shan, Y. Yue, and D. Wang, "Vision-based flexible leader-follower formation tracking of multiple nonholonomic mobile robots in unknown obstacle environments," *IEEE Transactions on Control Systems Technology*, pp. 1–9, 2019.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[5] J. S. Berrio, W. Zhou, J. Ward, S. Worrall, and E. Nebot, "Octree map based on sparse point cloud and heuristic probability distribution for labeled images," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3174–3181.

[6] Y. Yue, C. Yang, Y.Wang, P. G. C. N. Senarathne, J. Zhang, M. Wen, and D. Wang, "A multi-level fusion system for multi-robot 3d mapping using heterogeneous sensors," *IEEE Systems Journal*, 2019.

[7] Y. Yue, P. G. C. N. Senarathne, C. Yang, J. Zhang, M. Wen, and D. Wang, "Probabilistic fusion framework for collaborative robots 3d mapping," in *2018 21st International Conference on Information Fusion (FUSION)*, July 2018, pp. 488–491.

[8] Y. Yue, C. Yang, J. Zhang, M. Wen, Z. Wu, H. Zhang, and D. Wang, "Day and night collaborative dynamic mapping in unstructured environment based on multimodal sensors," in *Robotics and Automation (ICRA), 2020 IEEE International Conference on*, May 2020.

[9] J. Colorado, I. Mondragon, J. Rodriguez, and C. Castiblanco, "Geo-mapping and visual stitching to support landmine detection using a low-cost uav," *International Journal of Advanced Robotic Systems*, vol. 12, no. 9, p. 125, 2015.

[10] Y. Yue, D. Wang, P. G. C. N. Sena, and C. Yang, "Robust submap-based probabilistic inconsistency detection for multi-robot mapping," in *2017 European Conference on Mobile Robots (ECMR)*, Sept 2017, pp. 1–6.

[11] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert, "Initialization techniques for 3d slam: a survey on rotation estimation and its use in pose graph optimization," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 4597–4604.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[13] S. Yang, Y. Huang, and S. Scherer, "Semantic 3d occupancy mapping through efficient high order crfs," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 590–597.

[14] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[15] H. Yu, J. Moon, and B. Lee, "A variational observation model of 3d object for probabilistic semantic slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5866–5872.

[16] C. Yang, D. Wang, Y. Zeng, Y. Yue, and P. Siritanawan, "Knowledge-based multimodal information fusion for role recognition and situation assessment by using mobile robot," *Information Fusion*, vol. 50, pp. 126 – 138, 2019.

[17] I. A. Bârsan, P. Liu, M. Pollefeys, and A. Geiger, "Robust dense mapping for large-scale dynamic environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7510–7517.

[18] D. Kochanov, A. Ošep, J. Stückler, and B. Leibe, "Scene flow propagation for semantic mapping and object discovery in dynamic street scenes," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1785–1792.

[19] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-Robot Simultaneous Localization and Mapping: A Review," *Journal of Field Robotics*, vol. 33, no. 1, pp. 3–46, 2016.

[20] S. Saeed, P. Liam, T. Michael, and Seto, "Group mapping: A topological approach to map merging for multiple robots," *Robotics & Automation Magazine, IEEE*, vol. 21, no. 2, pp. 60–72, 2014.

[21] A. Gawel, T. Cieslewski, R. Dub, M. Bosse, and R. Siegwart, "Structure-based vision-laser matching," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 182–188.

[22] Y. Yue, D. Wang, P. Senarathne, and D. Moratuwage, "A hybrid probabilistic and point set registration approach for fusion of 3D occupancy grid maps," in *2016 IEEE International Conference on Systems, Man, and Cybernetics*, Oct 2016, pp. 1975–1980.

[23] C. Brand, M. J. Schu, and H. Hirsch, "Submap matching for stereo-vision based indoor/outdoor SLAM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2015, pp. 5670–5677.

[24] T. M. Bonanni, B. D. Corte, and G. Grisetti, "3-D Map Merging on Pose Graphs," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1031–1038, April 2017.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[26] J. Zhang, P. Siritanawan, Y. Yue, C. Yang, M. Wen, and D. Wang, "A two-step method for extrinsic calibration between a sparse 3d lidar and a thermal camera," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2018.

[27] Y. Yue, P. G. C. N. Senarathne, C. Yang, J. Zhang, M. Wen, and D. Wang, "Hierarchical probabilistic fusion framework for matching and merging of 3-d occupancy maps," *IEEE Sensors Journal*, vol. 18, no. 21, pp. 8933–8949, Nov 2018.

[28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3213–3223.

[29] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time," in *Robotics: Science and Systems Conference*, July 2014.