Collaborative semantic understanding and mapping framework for autonomous systems

Yue, Yufeng; Zhao, Chunyang; Wu, Zhenyu; Yang, Chule; Wang, Yuanzhe; Wang, Danwei

2020

Yue, Y., Zhao, C., Wu, Z., Yang, C., Wang, Y. & Wang, D. (2020). Collaborative semantic understanding and mapping framework for autonomous systems. IEEE/ASME Transactions On Mechatronics, 26(2), 978-989. https://dx.doi.org/10.1109/TMECH.2020.3015054

https://hdl.handle.net/10356/152421

https://doi.org/10.1109/TMECH.2020.3015054

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2020.3015054, IEEE/ASME Transactions on Mechatronics

Collaborative Semantic Understanding and Mapping Framework for Autonomous Systems

Yufeng Yue, Chunyang Zhao, Zhenyu Wu, Chule Yang, Yuanzhe Wang, and Danwei Wang

***Abstract*—Performing collaborative semantic mapping is a critical challenge for cooperative robots to enhance their comprehensive contextual understanding of the surroundings. This paper bridges the gap between the advances in collaborative geometry mapping that relies on pure geometry information fusion, and single robot semantic mapping that focuses on integrating continuous raw sensor data. 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 modelling of the hierarchical semantic map fusion framework and its mathematical derivation of its probability decomposition. At the single robot level, the semantic point cloud is obtained by combining information from heterogeneous sensors and used to generate local semantic maps. At the collaborative robots level, local maps are shared among robots for global semantic map fusion. Since the voxel correspondence is unknown between local maps, an Expectation-Maximization approach is proposed to estimate the hidden data association. Then, Bayesian rule is applied to perform semantic and occupancy probability update. Experimental results on the UAV (Unmanned Aerial Vehicle) and the UGV (Unmanned Ground Vehicle) platforms show the high quality of global semantic maps, demonstrating the accuracy and utility in practical missions.**

***Index Terms*—Collaborative Semantic Mapping, Semantic Segmentation, Information Fusion, Mobile Robots**

I. Introduction

As autonomous systems become more common in our daily lives, they are expected to interact with each other, share information, and execute tasks collaboratively. Collaborative robots have benefits of multi-perspective perception, high efficiency, and robustness to single robot failure, so they can perform more intelligent tasks such as human robot teaming and task-oriented robots formation. Due to limited sensing capabilities, each robot has only partial information of the surroundings. In such cases, the fundamental challenge is to deploy multiple robots that can perform collaborative semantic understanding and reconstruct the environment. Current 3D geometry mapping method only contains geometry information, which limits the application of robots in high level tasks. To enhance the perception capability of a group of robots, individual robots have to share local maps to generate a global representation, which is composed of geometry information and semantic

This research project is supported by National Research Foundation (NRF) Singapore, ST Engineering-NTU Corporate Lab under its NRF Corporate Lab@ University Scheme. *(Corresponding Author: Yufeng Yue)*

Yufeng Yue, Chunyang Zhao, Zhenyu Wu, Yuanzhe Wang and 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 (NIIDT), Academy of Military Science, Beijing, 100071, China

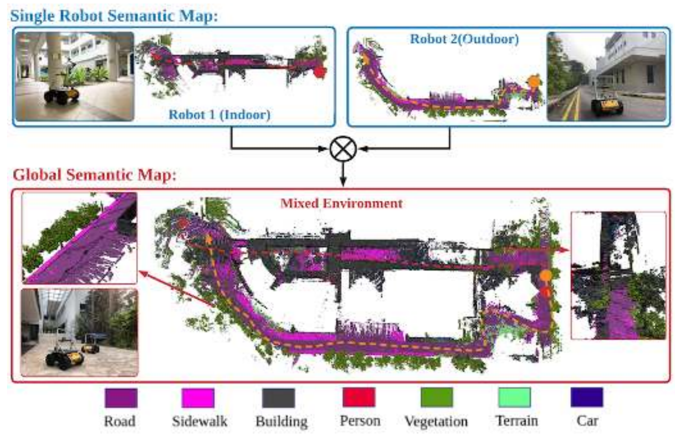


Fig. 1. The demonstration of collaborative semantic understanding and mapping in mixed indoor-outdoor environment. The red trajectory was followed by robot 1, while the orange one by robot 2. Top: local semantic maps generated by two robots. Bottom: collaboratively generated global semantic map.

context. Till now, the problem of collaboratively generating a consistent global semantic map using multiple robots remains open. This paper takes one step forward by focusing on collaborative probabilistic semantic mapping problem.

Single robot semantic mapping serves as the basic module for collaborative semantic mapping [1]. In recent years, multi-modal sensor-based [2] and Octree-based [3] algorithms have been proposed for single robot semantic mapping. However, there is a fundamental difference between single robot mapping and collaborative mapping because they work on different data levels. Single robot mapping aims to fuse the input continuous raw sensor scans into a local map, while collaborative mapping takes local maps generated by different robots as the input. The data heterogeneity makes a single-layer structure unqualified for collaborative mapping. In addition, it is not efficient to transmit raw sensor data between robots due to the bandwidth limitation. Therefore, the first key challenge is how to design a hierarchical collaborative mapping framework that is able to perform collaborative semantic perception efficiently under communication constraints.

Various approaches have been proposed to perform collaborative geometry mapping, such as estimating relative localization [4], accounting for transformation uncertainty [5], or updating maps in the global perspective [6]. However, existing collaborative geometry mapping methods focus on fusing geometry information such as planes, lines, and points. In the process of geometry information fusion, they simply fuse the nearest neighboring voxels without considering the disparity of semantic information. Collaborative semantic mapping differs from them because the data association between semantic voxels is unknown, which introduces another layer of difficulty. In summary, the second key challenge is how to estimate the hidden data association between semantic voxels, update and integrate probabilistic information into a consistent global semantic map.

This work is motivated by the fact that comprehensive analysis, modelling and implementation of collaborative semantic mapping methods have not been studied in-depth. This paper bridges the gap between collaborative geometry mapping that relies on pure geometry information fusion, and single robot semantic mapping that focuses on integrating continuous raw sensor data. The key novelty of this work is the modelling of the hierarchical semantic map fusion framework and the mathematical derivation of its probability decomposition. The main contributions are listed as follows:

- A hierarchical collaborative probabilistic semantic mapping framework is proposed and formulated in both single-robot and collaborative robots levels.
- An Expectation-Maximization (EM) algorithm is proposed to estimate the hidden data association between voxels in local maps, where the Bayesian rule is applied to perform semantic and occupancy probability updates.
- The proposed framework is validated in various experimental scenarios, demonstrating its accuracy and utility in practical tasks.

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

In this section, an extensive survey of existing approaches related to collaborative semantic mapping is provided. Among these solutions, semantic segmentation, single robot semantic mapping and collaborative geometry mapping are the most relevant domains.

A. Semantic Segmentation

Interpreting the scene is important for the robot to interact with the environment. The purpose of semantic segmentation is to assign a label to each pixel of an image. In recent years, convolutional neural networks (CNNs) have become the mainstream in computer vision tasks such as image classification and segmentation. For semantic segmentation tasks, the work in [7] presents a novel fully connected network (FCN) architecture Segnet, where the decoder up-samples its lower resolution input as the feature map. Refinenet [8] is proposed to exploit all the information for high-resolution image prediction. In Deeplab [9], multiple parallel atrous convolutions with different rates are involved to process the feature map. In this paper, Deeplab is employed to process the image data perceived by mobile robots.

B. Single Robot Semantic Mapping

To realize single robot semantic mapping, each robot needs to localize itself accurately in the environment. Up to now, the simultaneous localization and mapping (SLAM) for single robot is considered as a well-studied problem. Current approaches mainly consists of Lidar SLAM [10], visual SLAM [11] and graph SLAM [12]. However, they only restores the geometry information of the surroundings.

With the rapid development of deep learning, semantic mapping has attracted a lot of attention. A comprehensive review for single robot semantic mapping is summarized in [1]. The authors in [13] propose a hierarchical framework to combine spatial information and geometric maps in 2D environments. Then, the work in [14] propose an incremental semantic 3D mapping system for large-scale environments using the scrolling occupancy grid. To improve the efficiency, the authors in [3] proposes an octree based multi-label 3D semantic mapping algorithm. Since semantic information is also important for navigation [15], the authors in [16] use the semantic map to provide a hierarchical navigation solution. More recently, the authors in [17] applies object-level entities to construct semantic maps and integrate them into the semantic SLAM framework. The authors in [18] integrates semantic mapping with simultaneous object detection and localization. To perform mapping in dynamic environments, the authors in [19] propose a dense mapping algorithm based on stereo camera. Regarding the reconstruction of moving objects, the work in [20] incrementally fuses sensor observations into a consistent semantic map. The aforementioned approaches promote the development of single robot semantic mapping. However, single robot mapping mainly fuses incoming raw sensor data, while collaborative mapping takes local maps as input. This inherent difference prevents existing algorithms from being directly applied to collaborative semantic mapping, therefore a novel framework and theoretical formula are needed.

C. Collaborative Geometry Map Fusion

The coordination between multiple robots allows them to perform difficult tasks more efficiently and reliably. Distributed multi-robot coordination brings the challenges of data sharing [5], relative localization [21], and communication management [22]. The key challenge is to fuse the maps generated by individual robots (i.e., local maps) into a globally consistent map (i.e., a global map).

In order to balance the requirements of limited communication and detailed 3D mapping, global mapping requires careful selection of map types based on actual conditions [23]. Data fusion between multiple robots can be grouped into three different types, raw sensor data [24], volumetric maps [25] and topological maps [23]. When a robot receives local maps from neighboring robots, the key issue is to integrate the probabilistic information from these maps into a global consistent map. In most of the previous methods, the final maps are generated by stitching the overlap area [26] or averaging the geometry occupancy probability on the voxel-wise level [27]. To fuse semantic maps, the updating of semantic probability should also be considered. Collaborative semantic mapping differs from them because the data association between corresponding semantic voxels is unknown. Therefore, there is a strong need

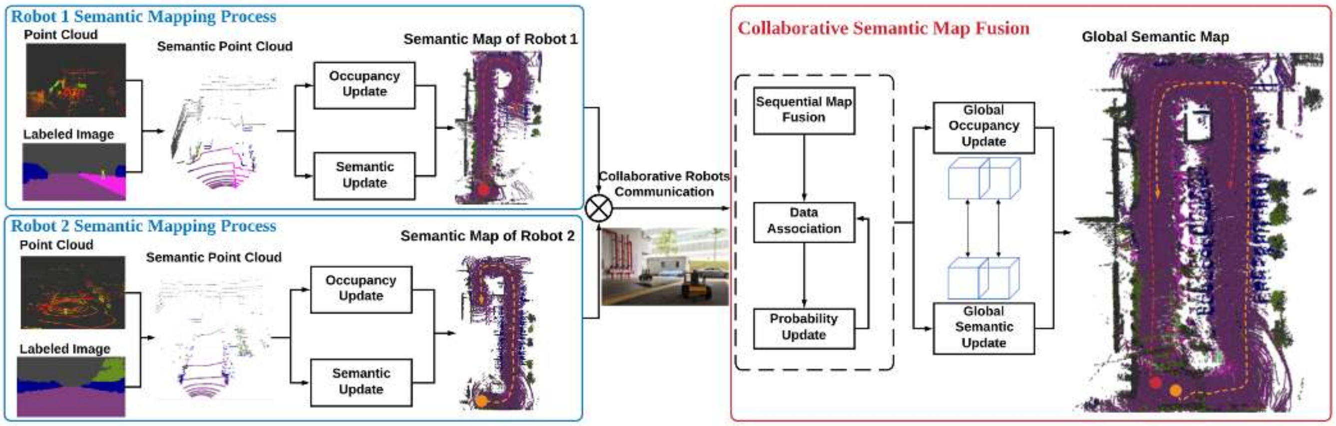


Fig. 2. The framework of hierarchical collaborative semantic understanding and mapping. The red trajectory was followed by robot 1, while the orange one by robot 2.

for a strategy to estimate hidden data associations and update the probability correctly.

III. System Framework

In this section, the system architecture for collaborative semantic mapping is presented, where the problem is formulated to provide a theoretical basis for the system.

A. The Framework of Hierarchical Semantic 3D Mapping

The objective of this paper is to develop a hierarchical framework for collaborative semantic mapping that enables a group of robots to generate a consistent global semantic map. 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*. Since the collaborative robot system operates in a distributed configuration, the organization of this paper follows a hierarchical structure from the single robot level to the multi-robot level.

For single robot level, the heterogeneous sensors carried by each robot are calibrated and integrated. The robots generate semantic point clouds based on image semantic segmentation and sensor fusion output. By applying the Bayesian rule, semantic label probability and occupancy probability are updated to generate a single-robot semantic map.

For multi-robot level, each robot communicates with neighboring robots to share local 3D semantic maps. The Expectation-Maximization (EM) approach is applied to estimate the hidden data association between voxels. Then, the local semantic maps are fused into a global consistent semantic map. The notational symbols used in this paper are defined in Table I.

B. Centralized Problem Formulation

Considering a group of robots moving through an unknown environment and attempting 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 trajectories $x_{1:t}^{(\mathcal{R})}$.*

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

TABLE I
The main notations used throughout the paper.

Symbol	Description
	Single Robot Level:
r	A single robot is denoted as r
$I_t^{(r)}$	Camera observation of robot r at time t
$L_t^{(r)}$	3D laser observation of robot r at time t
$L_{st}^{(r)}$	Semantic labels correspond to 3D laser points
$x_t^{(r)}$	Pose of robot r at time t
$m_t^{(r)}$	Local semantic map generated by robot r at time t
	Multi-Robot Level:
$\mathcal{R}$	Set of all robots $\mathcal{R} \triangleq \{r\}_{1:R}$, R is number of robots
$\mathcal{R}_r$	Set of neighboring robots $\mathcal{R}_r \triangleq \{r_n\}_{1:R_n}$ in communication with r, where the number is R_n
M_r	The global map generated by robot r
T_{r,r_n}	Relative position between robot r and r_n

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, which 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 and the semantic label o_i. The label o_i comes from a finite set of discrete class labels: $S = \{1, 2, \cdots, k, \cdots\}$. For the input, 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 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 with limited bandwidth. Therefore, the paper adopts the distributed setting, and each robot builds its own global semantic map.

C. Distributed Hierarchical Definition

In constrained environments, 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 communication node. In order to improve network efficiency and robustness, only local semantic maps are shared among the group of robots. The collaborative robots level mainly estimates the global semantic map given the output 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 the single robot level, multimodal semantic information fusion model serves as the input for each robot. Prior to single robot semantic mapping, the multimodal information fusion algorithm is applied to generate semantic point cloud. The process of multimodal environmental perception is presented in IV-A.

Given the input of semantic point cloud, each robot will estimate its local semantic map $m_t^{(r)}$. For semantic mapping, the occupancy probability and semantic label probability are updated simultaneously. Based on conditional independence, each voxel in the semantic map is updated separately with occupancy update model and semantic update model. The details of single robot semantic mapping are presented in IV-B.

The benefits of introducing the single robot level comes at the following advantages, firstly, the size of the local semantic map is greatly compressed compared to the raw sensor data (point cloud and image), which significantly reduces the communication burden; Besides, the factorization also gives the flexibility of choosing various SLAM and semantic segmentation algorithms.

Collaborative Robots Level Definition: *The objective of 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* $\mathcal{R}_r$.

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

At the collaborative robots level, it is assumed that each robot r has estimated its semantic map $m_t^{(r)}$ based on local observations. However, the perception ability and operation area of a single robot are limited. The comprehensive understanding of the environment should be obtained by integrating information from neighboring robots. Hence, robot r receives the local map $m_t^{(r_n)}$ from the nearby robots $r_n \in \mathcal{R}_r$, where $\mathcal{R} = \{r \cup \mathcal{R}_r\}$ if the communication covers all robots.

Under limited communication, robot r receives neighboring maps in a certain time interval. Hence, the map fusion is performed serially. Given the received local maps, 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 (see IV-C).

IV. Collaborative Semantic 3D Mapping

This section presents a hierarchical collaborative semantic mapping framework, which is divided into three subsections: multimodal semantic information fusion, single robot semantic mapping and collaborative semantic map fusion.

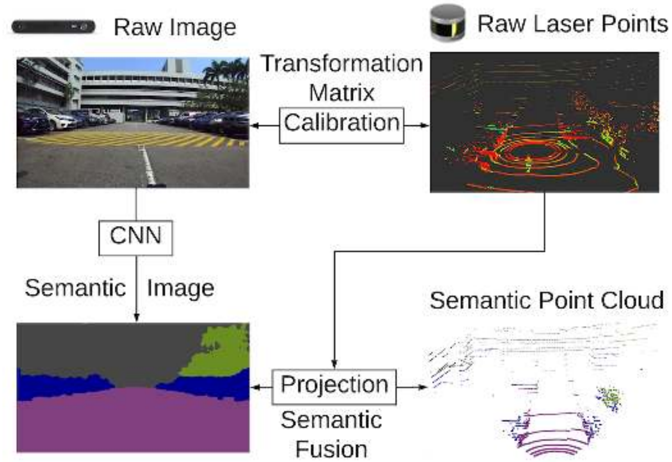


Fig. 3. Multimodal semantic information fusion model.

A. Multimodal Semantic Information Fusion

The goal of multi-modal semantic information fusion is to assign a semantic label for each point in a 3D point cloud. In this work, we utilize the 3D LiDAR and the visual camera as the main sensors for semantic mapping. At time t, to obtain a semantic point cloud, the semantic image I_{s_t} is generated by passing the raw image I_t through the Deeplab model [7]. Deeplab can output 19 semantic classes, including road, tree, building, and pedestrian etc. Then, we assign the semantic classes from the semantic image to the point cloud by utilizing the calibration parameters and projection equation: first, the 3D point cloud L_t is projected onto the semantic image I_{s_t} with the projection equation (4). Due to the limited field of view of the camera, only 3D points that fall into the semantic image will receive semantic information. Here, L_t^i is the 3D point in the point cloud $L_t^i \in L_t$, T_l^c is the extrinsic parameter between the 3D LiDAR and the camera, K_c is the intrinsic parameter of the camera, l_t^i is the projected point [28]. After projection, the projected point l_t^i is overlapped with the 2D pixel $p_{s_t}^i$ in the semantic image I_{s_t} ($p_{s_t}^i \in \mathrm{I}_{s_t}$). Therefore, the semantic label of pixel $p_{s_t}^i$ can be assigned to point l_t^i and L_t^i. Thus, each 3D point L_t^i receives the semantic information. Subsequently, a semantic point cloud L_{s_t} can be obtained.

$$l_t^i = K_c T_l^c L_t^i \tag{4}$$

The generated semantic point cloud consists of $L_{1:t}^{(r)}$ and $L_{s_{1:t}}^{(r)}$, where $L_{1:t}^{(r)}$ denotes 3D geometry coordinate and $L_{s_{1:t}}^{(r)}$ represents the corresponding $19D$ semantic labels, the subscript $1:t$ means from time 1 to t, the superscript (r) denotes the robot number r. In the semantic mapping process, raw images $I_{1:t}^r$ can not be used as the input. Therefore, the semantic point cloud will serve as the input. Then, (2) can be rewritten as:

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

B. Single Robot Semantic Mapping

Single robot localization problem is the basis for semantic mapping, which has been well studied and many efficient approaches have been developed. Any SLAM methods can be applied to estimate the odometry $x_{1:t}$. This paper focuses on collaborative mapping and we assume that the local maps are consistent. As the laser scan is with respect to its local sensor frame, the semantic point cloud is converted into the map frame using odometry $x_{1:t}$, where $\hat{L}_t = x_t * L_t$ denotes the transformed 3D geometry coordinate, $\hat{L}_{st} = x_t * L_{st}$ denotes the transformed semantic labels, then we have

$$p(m_t^{(r)}|L_{s1:t}^{(r)}, L_{1:t}^{(r)}, x_{1:t}^{(r)}) = p(m_t^{(r)}|\hat{L}_{1:t}^{(r)}, \hat{L}_{s1:t}^{(r)}) \tag{6}$$

As a result, the probability distribution of map m_r is estimated given observation $\hat{L}_{1:t}^{(r)}$ and $\hat{L}_{s1:t}^{(r)}$ as shown in (6). In the following content, the denotation of $m_t^{(r)}$ is simplified as m_r. For the local semantic map, we define $m_r = \{m_r^i\}_{i=1}^{N_{m_r}}$, where N_{m_r} is the number of voxels in the map. 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:

$$p(m_r|\hat{L}_{s1:t}^{(r)}, \hat{L}_{1:t}^{(r)}) = \prod_{i=1}^{N_{m_r}} p(m_r^i|\hat{L}_{s1:t}^{(r)}, \hat{L}_{1:t}^{(r)}) \tag{7}$$

Then, the probability of a leaf node m_r^i is estimated by updating semantic label probability and occupancy 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 $p(o_r^i|\hat{L}_{s1:t}^{(r)})$, then (7) is rewritten as (8).

$$\prod_{i=1}^{N_{m_r}} p(v_r^i, o_r^i|\hat{L}_{s1:t}^{(r)}, \hat{L}_{1:t}^{(r)}) = \prod_{i=1}^{N_{m_r}} \underbrace{p(v_r^i|\hat{L}_{1:t}^{(r)})}_{\text{Occupancy Update}} \cdot \underbrace{p(o_r^i|\hat{L}_{s1:t}^{(r)})}_{\text{Semantic Update}} \tag{8}$$

1) Voxel Occupancy Update: The occupancy probability is recursively updated given the incoming 3D point cloud $\hat{L}_{1:t}^{(r)}$. Based on Bayesian rule, the occupancy update model is expanded and updated in (9). The updating of (8) depends on the current observation $p(v_r^i|\hat{L}_t^{(r)})$, the previous estimation $p(v_r^i|\hat{L}_{1:t-1}^{(r)})$ and the prior probability $p(v_r^i)$. $p(v_r^i)$ denotes the initial occupancy probability of each voxel and is set as 0.5. The details of deriving (9) with Bayesian rule can be found in [29].

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

2) Semantic Label Update: The semantic probability is recursively updated given the incoming set of 19 semantic labels $\hat{L}_{s1:t}^{(r)} = \{\hat{L}_{s1:t}^{(r)}(k)\}_{k=1}^{19}$. Then, $p(o_r^i|\hat{L}_{s1:t}^{(r)})$ denotes the probability of updated label class (k=1:19). Based on Bayesian rule, the semantic update model is expanded and updated in (10). The updating of (10) depends on the current observation $p(o_r^i|\hat{L}_{st}^{(r)})$, the previous estimation $p(o_r^i|\hat{L}_{s1:t-1}^{(r)})$ and the prior probability $p(o_r^i)$. $p(o_r^i)$ denotes the initial semantic probability of each voxel and is set as $\frac{1}{19}$. The details of deriving (10) with Bayesian rule can be found in [29].

$$p(o_r^i|\hat{L}_{s1:t}^{(r)}) = [1 + \frac{1-p(o_r^i|\hat{L}_{st}^{(r)})}{p(o_r^i|\hat{L}_{st}^{(r)})} \frac{1-p(o_r^i|\hat{L}_{s1:t-1}^{(r)})}{p(o_r^i|\hat{L}_{s1:t-1}^{(r)})} \frac{p(o_r^i)}{1-p(o_r^i)}]^{-1} \tag{10}$$

After fusion, the class that corresponds to maximum probability is assigned 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, maps are generated and transmitted sequentially in a time interval. This results in robot r receiving its neighboring local 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. (11). Initially, the relative transformation matrix T_{r,r_n} between m_r and m_{r_n} is unknown. To estimate T_{r,r_n}, this paper applies the map matching algorithm proposed in [30]. Then, the neighboring robot map m_{r_n} is transformed to coordinate frame of robot r by transformation function $\Phi(T_{r,r_n}, m_{r_n})$, which is simplified as $\Phi(m_{r_n})$.

$$p(M_r|m_r, \Phi_{r_n \in \mathcal{R}_r}(m_{r_n})) = \prod_{r_n=1}^{N_r} p(M_r^{\pi(1:r_n)}|M_{r_n}^{\pi(1:r_n-1)}, \Phi(m_{\pi(r_n)})) \tag{11}$$

where $M_r^{\pi(1:r_n)}$ is the global map by fusing $M_r^{\pi(1:r_n-1)}$ with latest incoming map $m_{\pi(r_n)}$. The final global fused map M_r of robot r can be retrieved after the end of this serial process, hence $M_r = M_r^{\pi(1,Nr)}$.

2) Global Map Occupancy and Label Update: The key difference between local and global semantic map update comes from two aspects. 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, 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 classes. To show the difference, an example is presented in Fig. 5. In the left image, the grey color denotes geometry map. As can be seen, the geometry map fusion is based on closest neighborhood distance search, which ignores the object attributes. In contrast, the right image shows an example of semantic map fusion, where blue voxels denote the car and green voxels represent the grass. In this case, semantic data association should reject the wrong correspondences and accept correct pairs by incorporating semantic information.

For semantic map fusion, it is vital to come up with a strategy to consider the dissimilarities and fuse them into the global semantic map (Fig. 4(d)). The following two subsections will detail the hidden data association estimation and global information fusion.

For simplification, we denote the latest incoming neighboring robot map as m_{r_n} and current robot map as m_r.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2020.3015054, IEEE/ASME Transactions on Mechatronics

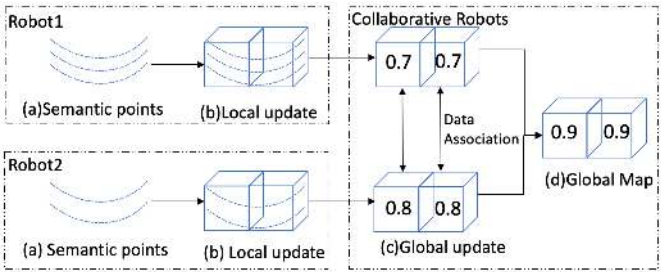
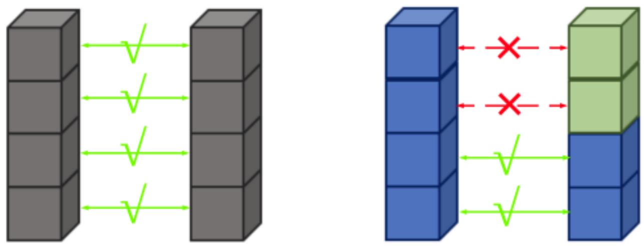


Fig. 4. An example of local and global semantic map update process. The value represents semantic probability, indicating the semantic update process.



(a) Geometry map fusion

(b) Semantic map fusion

Fig. 5. An example of geometry map fusion process in Fig. 5a, it only considers the geometry information and simply stitch the map. While in Fig. 5b, semantic information is considered to establish correct data associations. The red cross rejects the wrong data association.

Therefore, the probability distribution $p(M_r|m_r, \Phi_{r_n \in \mathcal{R}_r}(m_{r_n}))$ in (11) is simplified as $p(M_r|m_r, m_{r_n})$. Since the voxel-wise correspondences are usually assumed to be independent, the probability distribution of $p(M_r|m_r, m_{r_n})$ can be factorized into (12).

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

To estimate the hidden data association between voxel m_r^i and voxel $m_{r_n}^j$, a binary variable $d_{i,j}$ is introduced to represent the data association between corresponding voxels m_r^i and $m_{r_n}^j$. We have $d_{i,j} = 1$ if m_r^i corresponds to $m_{r_n}^j$ and $d_{i,j} = 0$ otherwise. Based on total probability theory, (12) is rewritten in the form of (13).

$$p(M_r|m_r, m_{r_n}) = \prod_{i=1}^{n} \sum_{j} p(M_r^k, d_{i,j}|m_r^i, m_{r_n}^j) \tag{13}$$

To solve (13), Expectation-Maximization (EM) is applied, which is an algorithm well-suited for models containing latent variables. The mathematical derivation of EM can be found in [31]. There are two steps of EM: the E-step efficiently estimates the hidden variables by evaluating the expectation, while M-step updates the global map probability given the corresponding voxel pairs (see (14)).

$$p(M_r^k, d_{i,j}|m_r^i, m_{r_n}^j) = \underbrace{p(d_{i,j}|m_r^i, m_{r_n}^j)}_{\text{E step:data association}} \cdot \underbrace{p(M_r^k|d_{i,j}, m_r^i, m_{r_n}^j)}_{\text{M step:probability update}} \tag{14}$$

E-Step: The E-step establishes the correspondence by calculating the minimum distance metric. 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 m_i^x, m_i^y, m_i^z ($3D$), occupancy probability v_{m_i} ($1D$) and semantic label o_{m_i} ($1D$). The corresponding voxel is calculated by finding the nearest neighborhood as formulated in (15). Once the nearest neighborhood within a certain threshold is found, the data association is established between m_r^i and $m_{r_n}^j$ with $d_{i,j} = 1$. Otherwise, we let $d_{i,j} = 0$. An example of E-Step with semantic information is shown in Fig. 5b.

$$\begin{aligned} p(d_{i,j}|m_r^i, m_{r_n}^j) &= d_e(m_r^i, m_{r_n}^j) + d_v(v_{m_r^i}, v_{m_{r_n}^j}) + o_v(o_{m_r^i}, o_{m_{r_n}^j}) \\ &= ||m_r^i - m_{r_n}^j||^2 + ||v_{m_r^i} - v_{m_s^j}||^2 + ||o_{m_r^i} - o_{m_s^j}||^2 \end{aligned} \tag{15}$$

M-Step: Given the data association $d_{i,j}$ and corresponding voxel pairs m_r^i, $m_{r_n}^j$, the next step is to fuse corresponding voxels into the global voxel M_r^k. (16) is the M-step of updating the probability, which is defined in (14). Since each voxel contains occupancy probability and semantic label probability, (16) is rewritten into (17). Considering the occupancy update and semantic update are updated independently, (17) is factorized into (18) as a global occupancy update model and a global semantic update model.

$$p(M_r^k|d_{i,j}, m_r^i, m_{r_n}^j) \tag{16}$$

$$= p(v(M_r^k), o(M_r^k)|d_{i,j}, v(m_r^i), v(m_{r_n}^j), o(m_r^i), o(m_{r_n}^j)) \tag{17}$$

$$= \underbrace{p(v(M_r^k)|d_{i,j}, v(m_r^i), v(m_{r_n}^j))}_{\text{Global Occupancy Update}} \cdot \underbrace{p(o(M_r^k)|d_{i,j}, o(m_r^i), o(m_{r_n}^j))}_{\text{Global Semantic Update}} \tag{18}$$

In single robot occupancy updating process (9), the input is a set of raw semantic points. However, the input in collaborative robots level is the probability value of local semantic map. Each voxel can be regraded as a Gaussian distribution. In this case, the fusion at collaborative robots level is the integration of two Gaussian distributions. The initial semantic probability of each global map voxel $p(v(M_r^k))$ is 0.5. Then, the global occupancy update is formulated based on the Bayesian rule, where details are provided in [29].

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

For the global semantic label probability update, the input is 19 class probabilities of each voxel from the single robot map. The initial semantic probability of each global map voxel $p(o(M_r^k))$ is $\frac{1}{19}$ (see (20)), where details are provided in [29].

$$\begin{aligned} &p(o(M_r^k)|d_{i,j}, o(m_r^i), o(m_{r_n}^j)) \\ &= [1 + \frac{1 - p(o(m_{r_n}^j))}{p(o(m_{r_n}^j))} \frac{1 - p(o(m_r^i))}{p(o(m_r^i))} \frac{p(o(M_r^k))}{1 - p(o(M_r^k))}]^{-1} \end{aligned} \tag{20}$$

After fusion, the probability values for all 19 labels are updated. The semantic probability of the most likely class $p(o(M_r^k), max)$ for each voxel M_r^k is computed as follows:

$$p(o(M_r^k), max) = \arg\max[p(o(M_r^k), 1), \ldots, p(o(M_r^k), 19)] \tag{21}$$

As a result, the class that corresponds to maximum probability is assigned to the class label of the voxel.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2020.3015054, IEEE/ASME Transactions on Mechatronics

TABLE II
COMPARISON OF CORE FUNCTIONALITY WITH EXISTING APPROACHES.

Algorithms	Category	Multimodal Perception	Geometry Mapping	Semantic Mapping	Collaborative Fusion	Distributed Communication
Yang et al. [14]	Single Robot	×	✓	✓	×	×
Berrio et al. [3]	Single Robot	✓	✓	✓	×	×
Saeedi et al. [23]	Multiple Robots	×	✓	×	✓	×
Jessup et al. [27]	Multiple Robots	×	✓	×	✓	×
This work	Multiple Robots	✓	✓	✓	✓	✓

(a) Open Carpark (b) Mixed Environment (c) UAV-UGV Mapping

Fig. 6. The experiment setting up: Open Carpark, Mixed Environment, UAV-UGV Mapping respectively.

V. EXPERIMENTAL RESULTS

The performance of the proposed collaborative semantic mapping framework is validated through extensive real-world experiments. Both qualitative and quantitative results are presented. The former is obtained by visualizing the fused map in various experimental settings. Then, the latter is expressed by showing average entropy, update process and size of data.

A. Evaluation Overview

1) Experimental Platform: Experiments are conducted by operating the UGV platforms in real environments. All algorithms are implemented on the ROS platform [32]. Two UGV platforms (Husky Clearpath) are equipped with the Intel Core i7-6700HQ CPU @ 2.60GHz CPU and the Nvidia GeForce RTX 2060 @ 6GB RAM GPU. The UGVs are also equipped with a 3D Velodyne LiDAR and a visual camera, where the sensors have already been calibrated. The UAV platform is designed by ourselves and is equipped with the ZED stereo camera with Intel NUC 6i7KYK @ 2.60GHz CPU. The communication between robots is established by long range Wi-Fi with limited bandwidth.

For the software packages, LOAM [10] is applied for UGV pose estimation. ORB-SLAM [11] is applied for UAV pose estimation. The Cityscapes dataset [33] is used to train the semantic segmentation model. HPFF [30] is implemented to estimate relative transformation between robots, and Octomap [34] is utilized for basic 3D geometry mapping with a resolution of 0.2m.

2) Experiment Setup: Experiments are conducted by operating UGV or UAV in three different scenarios in Nanyang Technological University (Fig. 6). All algorithms are executed on the ROS platform for multimodal semantic information fusion, single robot semantic mapping and collaborative semantic map fusion.

- **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.
- **UAV-UGV Semantic Mapping**: one Husky robot equipped with Velodyne 3D LiDAR and visual camera, one UAV equipped with ZED stereo camera.

3) Comparison Baseline: Most of the existing approaches either focus on single robot semantic mapping or collaborative geometry mapping. A direct qualitative comparison of our method with other approaches is conducted and summarized in Table II. The existing work can achieve part of the functions, but this paper is the only work that addresses the collaborative semantic mapping problem and realizes all functions. As no available work has addressed collaborative semantic mapping problem, we first demonstrate extensive qualitative results of how the global map is updated and how the local maps are fused. Then, we present quantitative results on average entropy, update process and size of data.

B. Open Carpark

To evaluate the multimodal semantic fusion algorithm for single robot semantic mapping, we first test our system in the Open Carpark. After processing the input data, semantic 3D reconstruction results can be obtained, which is presented in Fig. 7. The figure shows top view of the map and three close-up views for each time step. As shown in Fig. 7, the algorithm successfully integrates and updates the semantic information into 3D semantic map. More specifically, our approach can reconstruct the 3D map and recognize the classes of objects on the road with high accuracy.

The collaborative semantic mapping results is shown in Fig. 8. Two robots started their mission from nearby places and explored the environment by traversing two different paths. Fig. 8 shows the fusion of maps at start, middle and final of the mission. The left two columns show two local semantic maps. The right side is the fused global semantic map. For the overall process of collaborative mapping, it is shown that our system performs well in both single robot and multi-robot levels. For single robot level, the semantic map can be built by utilizing the raw sensor data perceived by each robot. For collaborative robots, the algorithm successfully combines the map information from two local maps to generate a consistent global map.

C. Mixed Indoor Outdoor

The mixed environment presented in Fig. 9 is a building in the university, where the indoor environment is the lobby, and the outdoor environment is the outside road. Robot **1** travelled

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2020.3015054, IEEE/ASME Transactions on Mechatronics

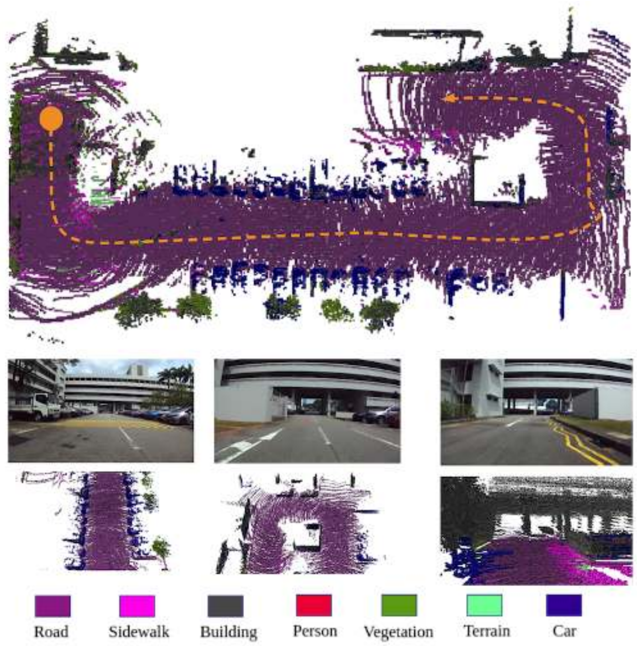
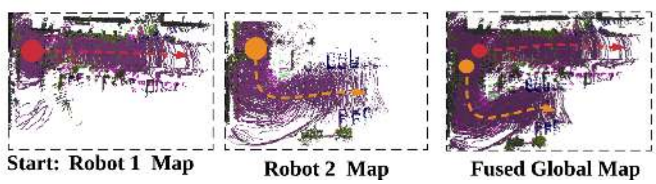
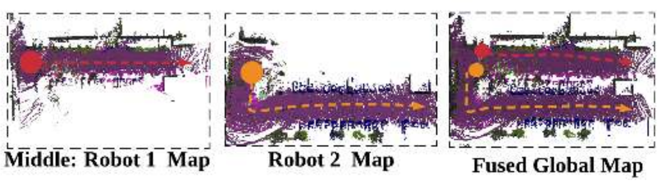


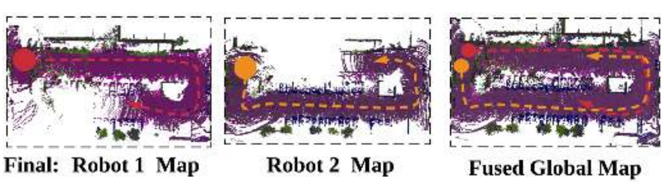
Fig. 7. Robot 2 semantic map in Open Carpark. Only the seven classes that appear in the scene are marked.



(a) t=0s, the initial position of the two robots. Two sparse local maps are successfully fused into a global map.



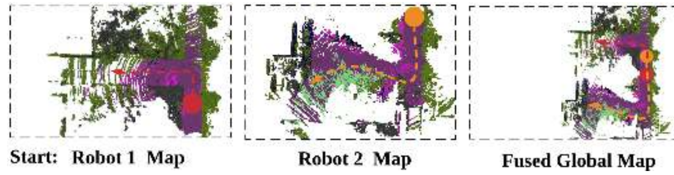
(b) Half of the mission. The local maps become denser, and global map are fused.



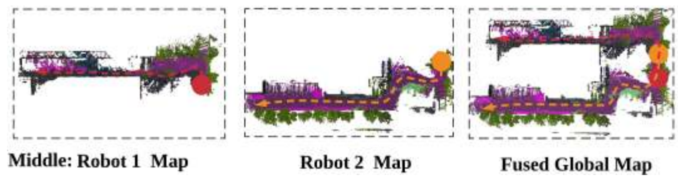
(c) End of the mission. The global map of the full environment is presented.

Fig. 8. Collaborative semantic mapping at the start, middle and final of the task in the open carpark.

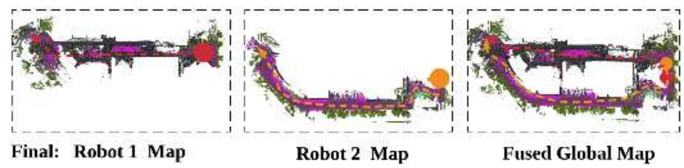
through the lobby in the building, and robot **2** moved along the road outside the building. The two robots first started nearby in the initial position, it can be seen that the two maps are successfully merged into a global map. Then, two robots encountered at the end position, and generated a global semantic map. Thanks to the probability update formulation, the overlapped area combines the advantages between the two maps to form a detailed global map.



(a) t=0s, the initial position of the two robots. Two sparse local maps are successfully fused into a global map.



(b) Half of the mission. The local maps become denser, and global map are fused.



(c) End of the mission. The global map of the full environment is presented.

Fig. 9. Collaborative semantic mapping at the start and final of the task in the mixed environment.

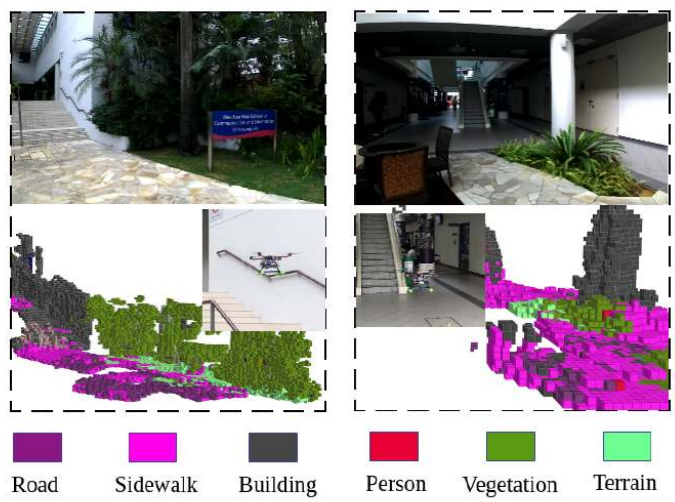


Fig. 10. The semantic map generated by UAV in two scenarios.

D. UAV-UGV Mapping

This part presents the result of UAV-UGV collaborative semantic mapping. It differs from the previous two experiments since heterogeneous robots and sensors are applied. In this environment, two robots started from a nearby place. The UGV could not cross the staircase, while the UAV directly flew over the staircase due to its high mobility (see Fig. 11). Fig. 10 demonstrates the semantic maps generated by UAV in two scenarios.

As presented in Fig. 11, UAV and UGV generated local semantic maps at the the end of the mission. Due to the specification of sensors, UAV generated a dense semantic map and successfully mapped the staircase. UGV captured most of the environmental information. Due to the limited mobility,

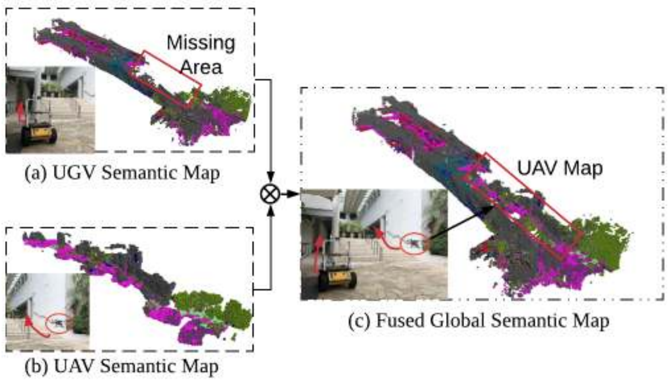


Fig. 11. The result of UAV-UGV semantic mapping. (a-b) Single robot SLAM is performed to generate the local semantic map. (c) The fused global semantic map.

TABLE III
GLOBAL MAP LABEL FUSION

	Robot1 (FP)		Robot2 (TP)		Global (TP)	
No	Label	Probability	Label	Probability	Label	Probability
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**

TABLE IV
GLOBAL MAP LABEL FUSION

	Robot1 (TP)		Robot2 (TP)		Global (TP)	
No	Label	Probability	Label	Probability	Label	Probability
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**

the UGV still failed to climb the staircase and map out the area as shown in Fig. 11. By combing the local semantic maps generated by UAV and UGV, a more detailed global semantic map is reconstructed. It shows the flexibility of the proposed algorithm in different robot platforms equipped with heterogeneous sensors.

E. Quantitative analysis

First, the average entropy is calculated to show the improved map uncertainty. We also examine the fusion accuracy in the global mapping process. Finally, the data size of the three experiments are summarized and compared.

1) Average Uncertainty: Table V summarizes the average entropy of the maps before and after collaborative mapping. In entropy theory, the higher the probability of the map, the lower the corresponding entropy. Therefore, we use average entropy to measure the uncertainty of the map. In Table V, the value presents the average entropy in different experiments. For the occupancy probability update, the entropy decreases after performing the proposed algorithm. For the label probability update, the entropy decreases as well. Then, the average entropy of overall probability update process decreases up to 10% in all experiments. Experiments show that the proposed map fusion strategy can effectively combine the probability information of two maps, thereby reducing the uncertainty of occupancy probability and label probability.

Fig. 12. Semantic probability update process of a voxel in global map.

2) Updating Process: Table III demonstrates the example of fusing mutual information from two local maps. In this scenario, the ground truth for the voxels from robot **1** and **2** should be Car. In the semantic mapping process, the voxel from robot **1** is assigned with the class of Road with a low probability of 0.638. However, it is a false positive (FP), and the ground truth label should be Car. The corresponding voxel in the robot **2** has a true positive (TP) class of Car with a high probability of 0.973. As can be seen from Table III, the voxel label in the global map is assigned to Car after fusion. This shows the superiority of the proposed algorithm for revising the false positive label after fusion, which increases the accuracy and robustness. Then, Fig. 12 shows a more detailed semantic probability update process of the global voxel in the global map, specifying the label and its corresponding probability. We can observe the increasing of the label Car (blue) and the decreasing in the remaining labels.

Table IV shows an example that the fusion process can enhance semantic probability of correctly assigned 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.

3) Data Size: To evaluate the efficiency of the collaborative systems, data management is worth studying. In the experiment, we recorded the accumulated data size of the entire process when executing the mission, as shown in Figure 13. The horizon axis is the duration of the mission and the unit is second. The vertical axis represents the data size, and is expressed in logarithmic of Bytes. The size of data generated by robot 1 is presented in three scenarios. It can be directly observed that the data volume of the raw image is nearly 10 GB, because the raw image records detailed information. Therefore, due to its large size, it is not suitable for sharing between robots. Then, the size of generated point cloud is also more than 1 GB, because the 3D Lidar records more than 100,000 points per second. The data volume of the semantic map is below 30MB, which is compact and suitable for communication between robots. In the experiment, we only share the generated 3D semantic map. In summary, sharing 3D semantic map will significantly reduce the communication burden and increase efficiency.

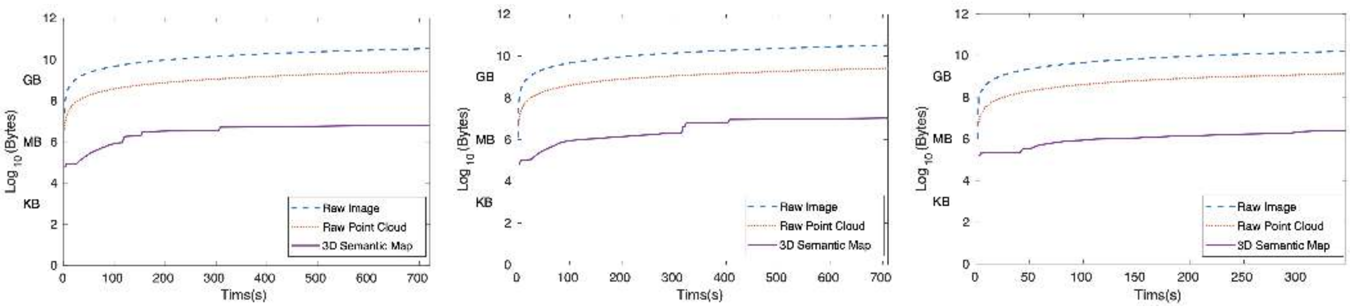
VI. CONCLUSION

This paper has established a hierarchical collaborative probabilistic semantic mapping framework. We have designed a new framework to provide the theoretical formulation and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMECH.2020.3015054, IEEE/ASME Transactions on Mechatronics

TABLE V
AVERAGE ENTROPY OF THE MAPS IN ALL EXPERIMENTS. (THE BEST PERFORMANCE IS DENOTED IN BOLD.)

Experiments	Occupancy Probability			Label Probability			Overall Probability		
	Map 1	Map 2	Global Map	Map 1	Map 2	Global Map	Map 1	Map 2	Global Map
Open Carpark	0.4030	0.3929	**0.3712**	0.3689	0.3653	**0.3429**	0.3439	0.3267	**0.3078**
WKW Lobby	0.3088	0.3285	**0.2849**	0.3239	0.3327	**0.2937**	0.2833	0.2930	**0.2715**
UAV-UGV	0.3291	0.3476	**0.3161**	0.3349	0.3527	**0.3229**	0.3098	0.3157	**0.2895**



(a) Size of Data in Open Carpark. (b) Size of Data in Mixed Environment. (c) Size of Data in UAV-UGV Mapping.

Fig. 13. The results of cumulative data size in all experiments.

system implementation for collaborative semantic mapping. In the single robot level, the semantic point cloud is obtained by combining information from heterogeneous sensors and is used to generate local semantic maps. To achieve collaborative semantic mapping, this paper has provided a theoretical basis for the global 3D semantic mapping. The results have shown that the proposed algorithm was able to establish the correct data association between voxels. More importantly, the fusion process is able to correct the false label and enhance the true label. The overall experimental results have presented high quality global semantic maps, which demonstrate the accuracy and utility of the framework. In summary, the proposed collaboration system provides a new perspective of sensing and reconstructing the environment, which is complementary to the individual robot perception and mapping.

In the future, relative localization between collaborative robots can be performed based on global semantic map, which will significantly improve the localization accuracy.

REFERENCES

[1] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robotics and Autonomous Systems*, vol. 66, pp. 86–103, 2015.

[2] J. Jeong, T. S. Yoon, and J. B. Park, "Multimodal sensor-based semantic 3d mapping for a large-scale environment," *Expert Systems with Applications*, vol. 105, pp. 1 – 10, 2018.

[3] 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.

[4] Y. Yue, C. Zhao, M. Wen, Z. Wu, and D. Wang, "Collaborative semantic perception and relative localization based on map matching," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[5] Y. Yue, C. Yang, Y. Wang, P. C. N. Senarathne, J. Zhang, M. Wen, and D. Wang, "A multilevel fusion system for multirobot 3-d mapping using heterogeneous sensors," *IEEE Systems Journal*, vol. 14, no. 1, pp. 1341–1352, 2019.

[6] 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.

[7] 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.

[8] 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.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

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

[11] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.

[12] H. Carrillo, P. Dames, V. Kumar, and J. A. Castellanos, "Autonomous robotic exploration using occupancy grid maps and graph slam based on shannon and renyi entropy," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, May 2015, pp. 487–494.

[13] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.-A. Fernandez-Madrigal, and J. González, "Multi-hierarchical semantic maps for mobile robotics," in *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2005, pp. 2278–2283.

[14] 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.

[15] H. Bayram and H. I. Bozma, "Multi-robot navigation with limited communication-deterministic vs game-theoretic networks," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 1825–1830.

[16] I. Kostavelis, K. Charalampous, A. Gasteratos, and J. K. Tsotsos, "Robot navigation via spatial and temporal coherent semantic maps," *Engineering Applications of Artificial Intelligence*, vol. 48, pp. 173–187, 2016.

[17] 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.

[18] Z. Zeng, Y. Zhou, O. C. Jenkins, and K. Desingh, "Semantic mapping with simultaneous object detection and localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 911–918.

[19] 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.

[20] 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.

[21] 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.

[22] A. Cunningham, V. Indelman, and F. Dellaert, "Ddf-sam 2.0: Consistent distributed smoothing and mapping," in *2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 5220–5227.

[23] 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.

[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] Y. Yue, C. Zhao, R. Li, C. Yang, J. Zhang, M. Wen, Y. Wang, and D. Wang, "A hierarchical framework for collaborative probabilistic semantic mapping," in *Robotics and Automation (ICRA), 2020 IEEE International Conference on*, May 2020.

[26] 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.

[27] J. Jessup, S. N. Givigi, and A. Beaulieu, "Robust and Efficient Multirobot 3-D Mapping Merging With Octree-Based Occupancy Grids," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–10, 2015.

[28] 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.

[29] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*. MIT press Cambridge, 2000, vol. 1.

[30] Y. Yue, C. Yang, 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.

[31] F. Dellaert, "The expectation maximization algorithm," Georgia Institute of Technology, Tech. Rep., 2002.

[32] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org

[33] M. Cordts, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3213–3223.

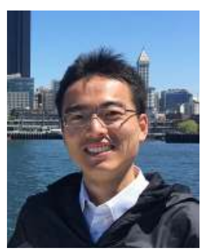
[34] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees," *Autonomous Robots*, 2013.

Yufeng Yue received the B.Eng. degree in automation from Beijing Institute of Technology, Beijing, China, in 2014, the Ph.D. degree from Nanyang Technological University, Singapore, in 2019. He is currently a Research Fellow with Advanced Robotics Lab, Nanyang Technological University, Singapore. He was a visiting scholar with University of California, Los Angeles in 2019. His research interests include perception, navigation and coordination for multi-robot system in complex environments.

Chunyang Zhao received the B.Eng. degree in control technology and instruments from Wuhan University, Wuhan, China, in 2019. He is currently a master student in Advanced Robotics Lab, Nanyang Technological University, Singapore. His research interests include machine learning, sensor fusion, localization and mapping for robot system.

Zhenyu Wu received his B.Eng. degree of Electrical Engineering and Automation from Wuhan University, China in 2016, and the M.Sc. degree of Electronics from Nanyang Technological University, Singapore in 2017. He is currently a Ph.D. student with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include mobile robot localization, multi-sensor fusion, and deep learning.

Chule Yang received his B.Eng. degree from Wuhan University, China, in 2014. He received his M.Sc. and Ph.D. degrees from Nanyang Technological University, Singapore, in 2015 and 2019, respectively. He is currently an assistant professor at National Innovation Institute of Defense Technology (NIIDT), Academy of Military Science. His research interests include swarm intelligence, collaborative semantic perception, and deep situation awareness for autonomous robots.



Yuanzhe Wang received the B.Eng. degree from the Southeast University, Nanjing, China, in 2010, the M.Eng. degree from the Beihang University, Beijing, China, in 2013, and the Ph.D. degree from the Nanyang Technological University, Singapore, in 2019. He has served as a Research Associate in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2018 to 2019, where he is currently a Research Fellow. His current research interests include robot motion planning and control, multi-robot coordination and human-robot interaction.

Danwei Wang is leading the autonomous mobile robotics research group. He received his Ph.D and MSE degrees from the University of Michigan, Ann Arbor in 1989 and 1984, respectively. He received his B.E degree from the South China University of Technology, China in 1982. Since 1989, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Currently, he is professor and the co-director of the ST Engineering NTU Corp Lab. He has served as general chairman, technical chairman and various positions in international conferences, such as ICRA and IROS conferences. His research interests include robotics, control theory and applications.