2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

HD-CCSOM: Hierarchical and Dense Collaborative Continuous Semantic Occupancy Mapping through Label Diffusion

Yinan Deng, Meiling Wang, Yi Yang, Yufeng Yue*

***Abstract*— The collaborative operation of multiple robots can make up for the shortcomings of a single robot, such as limited field of perception or sensor failure. Multi-robots collaborative semantic mapping can enhance their comprehensive contextual understanding of the environment. However, existing multi-robots collaborative semantic mapping algorithms mainly apply discrete occupancy map inference, and do not compensate for inconsistent labels of local maps caused by differences in robot perspectives, which leads to greatly reduced availability and accuracy of the final global map. To address the challenges of discontinuous maps and inconsistent semantic labels, this paper proposes a novel hierarchical and dense collaborative continuous semantic occupancy mapping algorithm (HD-CCSOM). This work decomposes and formulates robot collaborative continuous semantic occupancy mapping problem at two levels. At the single robot level, the multi-entropy kernel inference method smoothly processes the registered semantic point cloud and infers a local continuous semantic occupancy map for each robot. At the collaborative robots level, the local maps are fused into a global enhanced and consistent semantic map via the label diffusion method based on a graph model. The proposed algorithm has been validated on public datasets and in simulated and real scenes, demonstrating significant improvements in mapping accuracy and efficiency.**

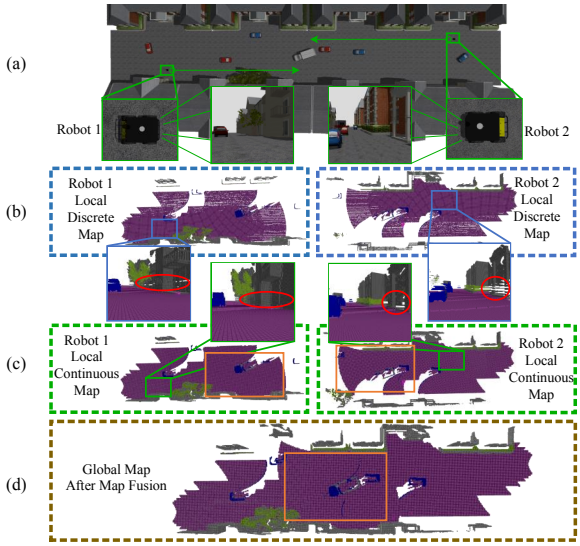

Fig. 1. A demonstration of HD-CCSOM. (a) The two robots move towards each other on either side of the road. (b) Discrete local maps with holes constructed by OctoMap [3]. (c) Dense local maps constructed by ours. (d) The global semantic map was obtained after map fusion..

I. Introduction

3D scene reconstruction plays a fundamental role in a series of robot tasks such as unknown environment exploration [1] and behavioral decision-making [2]. Current researches mainly focus on single robot mapping techniques [3]. Due to the limitation of perception ability, it is difficult for a single robot to comprehensively understand the environment from a limited perspective, and the reconstructed map may have a large deviation due to sensor noise. Comparatively, the collaborative operation of multi-robots can realize multi-perspective perception and improve mapping efficiency, which is a boon for time-critical tasks. In addition, semantic information can help robots to achieve a higher level of scene understanding [4], expanding the application of maps. Therefore, multi-robots collaborative semantic mapping has received increasing attention in recent years.

Multi-robots collaborative semantic mapping attempts to fuse different local semantic maps into a globally consistent semantic map. However, existing collaborative semantic mapping algorithms [5] assume complete independence between voxels, which brings difficulties in representing continuous object surfaces in the real world. These discrete mapping methods do not take into account the spatial association between sensor observations and voxels, resulting in a map with holes (See Fig. 1b). In addition, distributed robots are egocentric, so different robots may have different understandings even when exploring the same environment. Existing semantic map fusion methods [6] roughly fuse the labels of associated voxels without considering the credibility of local maps and the information of surrounding voxels, resulting in a downgraded global semantic map quality.

Therefore, this paper proposes a novel collaborative continuous semantic occupancy mapping algorithm HD-CCSOM that can reconstruct dense semantic maps and compensate for inconsistent semantic labels. Specifically, to reconstruct a continuous local semantic map at the single robot level, this paper adopts a multi-entropy kernel inference method, which can deal with sparse sensor data by inferring from surrounding observations. At the collaborative robots level, we formulate the semantic map fusion task as a graph-based label diffusion problem. Geometric context is leveraged to propagate labels from local maps to the global map, incorporating the credibility of local maps and the potential connections between voxels. All in all, we mathematically formulate the collaborative continuous semantic occupancy mapping problem and derive its probabilistic models.

This work is partly supported by the National Natural Science Foundation of China under Grant 62003039, 62173042, 61973034, U193203, and the CAST program under Grant No. YESS20200126. (Corresponding Author: Yufeng Yue, yueyufeng@bit.edu.cn)

All authors are with School of Automation, Beijing Institute of Technology, Beijing, 100081, China.

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9981756

II. Related Works

A. 3D Semantic Reconstruction

3D semantic reconstruction is extended from 3D geometric reconstruction [3], which restores not only the shape but also the class of the objects [7]. A typical representation is the semantic octree map [8], which embeds an octree into a hierarchical robust Markov Random Field. To optimize misclassified labels, Conditional Random Field (CRF) is applied to establish the association of local areas, such as 2D pixels [9] or 3D voxels [10]. However, it is based on the principle of discrete map inference and can only post-process the predicted labels, so that cannot fill the holes in the map. In addition, the aforementioned methods are addressing the problem of single robot semantic mapping, but do not consider multi-robots collaborative mapping.

B. Continuous Mapping

To reconstruct a smoother occupancy map, many approaches have attempted to relax the assumption that voxels are statistically independent. GPOM [11] introduces dependencies between points on a map by employing a modified Gaussian Process as a nonparametric Bayesian learning process. Hilbert map [12] can naturally capture spatial relationships between measurements by projecting the data into a Hilbert space, where a logistic regression classifier is learned. Recently, Bayesian Kernel Inference (BKI) [13] has emerged as a promising method to improve continuous mapping efficiency. Based on it, S-BKI [14] provides a unified probabilistic model to update occupancy and semantic probabilities. Unfortunately, these methods suffer from overfitting and high complexity, more importantly, they are not applied to collaborative robots.

C. Multi-Robots Collaborative Mapping

Multi-robots collaborative mapping has gained increasing attention in recent years. Group Mapping [15] proposes a strategy for 2D grid occupancy map fusion using a novel Probabilistic Generalized Voronoi Diagram (PGVD). Expectation-Maximization is applied in 3D collaborative geometric map fusion [16]. The first work to address collaborative semantic mapping is [17], bridging the gap between collaborative geometric mapping and single robot semantic mapping. Place Recognition [18], [19] can assist relative localization among multi-robots. To deal with novel observations, Jamieson et al. [20] applies unsupervised learning algorithms in multi-robots semantic mapping to reconstruct novel and unfamiliar environments. However, the above methods do not consider the credibility of different local maps and the association between voxels.

III. System Framework

A. The Systematic Framework

The objective of the collaborative continuous semantic occupancy mapping is to reconstruct a globally consistent semantic map for the explored environment using sensor observations from multiple robots. The framework of the proposed algorithm is depicted in Fig. 2, which can be divided into two levels. The main notational symbols are defined in Tab. I.

TABLE I
THE MAIN NOTATIONS USED THROUGHOUT THE PAPER

Symbol	Description
	Single Robot Level:
r_k	A single robot with index k
$m_t^{(r_k)}$	Local semantic map generated by robot r_k at time t
$v_i^{(r_k)}$	Voxel in local map $m_t^{(r_k)}$ with index i
	Collaborative Robots Level:
$\mathcal{R}$	A set of K robots $\mathcal{R} \triangleq \{r_k\}_{k=1}^{K}$
$\mathcal{M}_t^{(\mathcal{R})}$	Global semantic map generated by robots $\mathcal{R}$
$v_j^{(\mathcal{R})}$	Voxel in global map $\mathcal{M}_t^{(\mathcal{R})}$ with index j
	Mapping Details:
$I_t^{(r_k)}$	Camera observation of robot r_k at time t
$L_t^{(r_k)}$	3D LiDAR observation of robot at r_k time t
$O_t^{(r_k)}$	Pose of robot r_k at time t
$L_{s_t}^{(r_k)}$	Registered semantic point cloud of robot r_k at time t
$\mathcal{L}$	Semantic label sequence $\mathcal{L} \triangleq \{1, 2, \dots L\}$

For the single robot level, each robot performs the local continuous semantic occupancy mapping process independently. The registered semantic point cloud is obtained by sensor calibration and fusion, and then used to complete the continuous semantic occupancy map update through multi-entropy kernel inference.

For the collaborative robots level, robots will share their local maps with neighboring robots. By constructing a graph model, a label diffusion method is applied to fuse the local semantic maps into an accurate global semantic map.

B. Overall Problem Definition

Since the rate of information transfer among robots is limited, it is difficult to share the huge volume of raw sensor data in real time. Therefore, each robot independently generates a local semantic map as its own 'knowledge'. Only local semantic maps rather than raw sensor data will be shared among robots. When the robot receives the local semantic maps published by the neighboring robots, it fuses these maps into a global map to expand its 'knowledge'. Therefore, the problem can be split into two levels to achieve different tasks.

Single Robot Level Definition: *Given a robot r_k with its camera observation $I_{1:t}^{(r_k)}$, 3D LiDAR observation $L_{1:t}^{(r_k)}$ and trajectory $O_{1:t}^{(r_k)}$, the objective is to estimate its local semantic map $m_t^{(r_k)}$.*

$$p(m_t^{(r_k)} | I_{1:t}^{(r_k)}, L_{1:t}^{(r_k)}, O_{1:t}^{(r_k)}) \tag{1}$$

The introduction of the single robot level will greatly reduce the burden of raw sensor data transmission among robots. Distributed construction of local maps improves the efficiency of 3D reconstruction while providing flexibility in choosing SLAM methods and semantic segmentation algorithms for robots in different scenes.

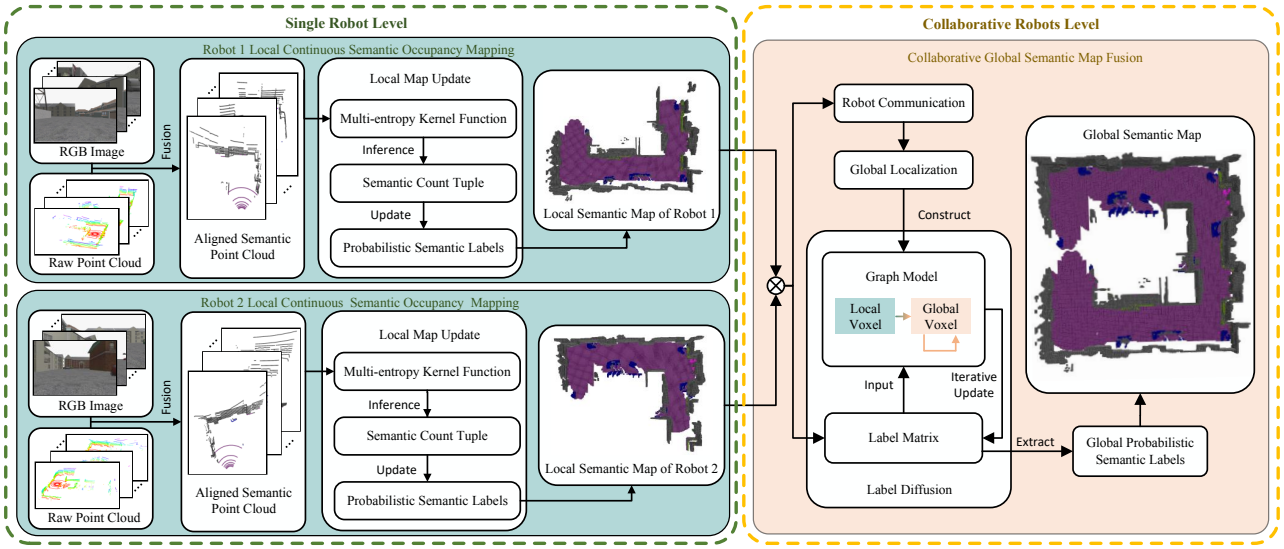

Fig. 2. The framework of hierarchical and dense collaborative continuous semantic occupancy mapping algorithm.

Collaborative Robots Level Definition: *Given local semantic maps* $\{m_t^{(r_k)}\}|_{r_k\in\mathcal{R}}$ *of all robots* $\mathcal{R}$*, the objective is to estimate the global semantic map* $\mathcal{M}_t^{(\mathcal{R})}$.

$$p(\mathcal{M}_t^{(\mathcal{R})}|\{m_t^{(r_k)}\}_{r_k\in\mathcal{R}}) \tag{2}$$

The output of the single robot level serves as the input of the collaborative robots level. The task at this level is to fuse the local semantic maps into a global semantic map. Aiming at the problem of inconsistent labels in local maps, a novel label diffusion method is applied to derive global map labels from multiple local maps.

IV. Collaborative Continuous Semantic Occupancy 3D Mapping

A. Single Robot Continuous Semantic Occupancy Mapping

In Subsection III-B, the task at the single robot level is defined by (1). At time t, the semantic segmentation network is activated to assign a semantic label to each pixel of the image $I_t^{(r_k)}$, which is then projected to the point cloud $L_t^{(r_k)}$ through sensor calibration and fusion [21]. By collecting points with labels and transforming them according to the pose $O_t^{(r_k)}$, the registered semantic point cloud $L_{s_t}^{(r_k)}$ is obtained.

For input, the semantic point cloud $L_{s_{1:t}}^{(r_k)} \triangleq \{p_n^{(r_k)}\}$ consists of a series of semantic points $p_n^{(r_k)}$. Each semantic point $p_n^{(r_k)}$ is represented by 3D coordinates and an associated one-hot encoded semantic label tuple $c_n^{(r_k)} = (c_n^{(r_k)1}, c_n^{(r_k)2}, \cdots c_n^{(r_k)L})$. Similarly, for output, the local semantic map $m_t^{(r_k)} \triangleq \{v_m^{(r_k)}\}$ consists of a set of voxels $v_m^{(r_k)}$. Each voxel $v_m^{(r_k)}$ is represented by 3D coordinates and an associated probabilistic semantic label tuple $\lambda_m^{(r_k)} = (\lambda_m^{(r_k)1}, \lambda_m^{(r_k)2}, \cdots \lambda_m^{(r_k)L})$, where $\sum_{l\in\mathcal{L}} \lambda_m^{(r_k)l} = 1$. Therefore, the problem of single robot continuous semantic occupancy mapping can be refined as:

$$p(m_t^{(r_k)}|L_{s1:t}^{(r_k)}) = \prod_n \prod_m p(\lambda_m^{(r_k)}|v_m^{(r_k)}, p_n^{(r_k)}, c_n^{(r_k)}) \tag{3}$$

Applying Bayesian rule, the posterior probability can be written as the product of the prior probability and the likelihood probability:

$$\begin{aligned} &p(\lambda_m^{(r_k)}|v_m^{(r_k)}, p_n^{(r_k)}, c_n^{(r_k)}) \\ &\propto p(\lambda_m^{(r_k)}) p(\, c_n^{r_k} \,|\, v_m^{(r_k)}, p_n^{(r_k)}, \lambda_m^{(r_k)}) \end{aligned} \tag{4}$$

where the likelihood probability satisfies the Categorical distribution $Cat(\lambda_m^{(r_k)1}, \lambda_m^{(r_k)2}, \cdots \lambda_m^{(r_k)L})$, while the prior and the posterior probability satisfy the Dirichlet distribution $Dir(L, \sigma_0^{(r_k)})$ and $Dir(L, \sigma_m^{(r_k)})$.

To reconstruct smooth surfaces of objects, a kernel function k operating in 3D space is introduced to continuously consider the association between semantic points $p_n^{(r_k)}$ and voxels $v_m^{(r_k)}$. A popular function k_0 [22] has the form as:

$$\begin{aligned} &k_0(v_m^{(r_k)}, p_n^{(r_k)}) \\ &= \mathop{\mathbf{I}}_{d<D} \varepsilon_0 \left[\frac{(2+\cos(2\pi\frac{d}{D}))(1-\frac{d}{D}))}{3} + \frac{\sin(2\pi\frac{d}{D})}{2\pi} \right] \end{aligned} \tag{5}$$

where $d = ||v_m^{(r_k)} - p_n^{(r_k)}||$, D is the fixed kernel length, and ε_0 is the scale factor. $\mathbf{I}$ represents the indicator function. To prevent overfitting and improve efficiency, the multi-entropy kernel function k_e is derived from k_0 in (6). Specifically, we introduce context entropy $\mathbf{E}$ to score the potential filled probability of voxels and filter out redundant voxels during continuous inference with a threshold $\mathbf{T}$. And the kernel length D_m is adaptively adjusted according to class entropy, which measures the overall uncertainty of the voxel $v_m^{(r_k)}$.

$$k_e(v_m^{(r_k)}, p_n^{(r_k)}) = \mathop{\mathbf{I}}_{\mathbf{E}<\mathbf{T}} k_0(v_m^{(r_k)}, p_n^{(r_k)})_{D\to D_m} \tag{6}$$

Substituting the likelihood probability after k_e expansion into (4) and simplifying it, the relationship between the two Dirichlet distribution parameters $\sigma_0^{(r_k)}$ and $\sigma_m^{(r_k)}$ is obtained:

$$\sigma_m^{(r_k)l} = \sigma_0^{(r_k)l} + \sum_n k_e(v_m^{(r_k)}, p_n{}^{(r_k)}) c_n^l \tag{7}$$

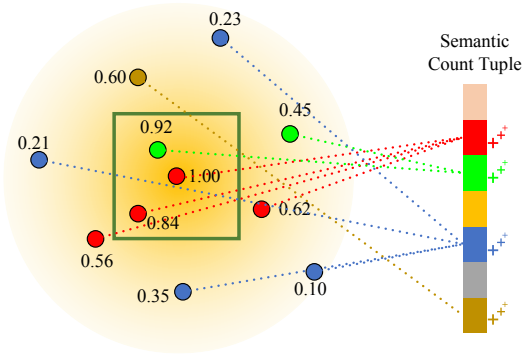

Fig. 3. An example of a semantic count tuple. Green box represent voxel $v_m^{(r_k)}$. The colored circles represent semantic points $p_n^{(r_k)}$ with different semantic labels $c_n^{(r_k)}$, and the numbers next to them represent the value of the multi-entropy kernel function. The semantic count tuple $\sigma_m^{(r_k)}$ absorbs these information incrementally.

As (7) defines, $\sigma_m^{(r_k)}$ is the weighted count of the semantic points, which is known as semantic count tuple. $\sigma_m^{(r_k)}$ counts not only the semantic points that fall into the current voxel $v_m^{(r_k)}$, but also the adjacent semantic points with the kernel function as the weight, thus realizing continuous semantic occupancy mapping, where an example is shown in Fig. 3. The associated semantic count tuple incrementally counts the semantic points within the voxel's kernel length.

The probabilistic label $\lambda_m^{(r_k)}$ of voxel $v_m^{(r_k)}$ in the local map $m_t^{(r_k)}$ is the closed-loop expected value of the posterior Dirichlet distribution $Dir(L, \sigma_m^{(r_k)})$:

$$\lambda_m^{(r_k)l} = \sigma_m^{(r_k)l} / \sum_{l' \in \mathcal{L}} \sigma_m^{(r_k)l'} \tag{8}$$

In general, the continuous semantic occupancy mapping of a single robot r_k is transformed into a mathematical solution. As the robot acquires new sensor observations, it only needs to incrementally solve (7) and (8) to complete the update of the local continuous semantic occupancy map $m_t^{(r_k)}$.

B. Collaborative Robots Global Semantic Map Fusion

At the collaborative robots level, it is assumed that all robots $r_k \in \mathcal{R}$ have completed the reconstruction of local semantic maps $m_t^{(r_k)}$. The problem at this level is defined in (2). The relative poses between local maps can be estimated by Multi-SLAM algorithms such as COSEM [23] or SectionKey [24]. For the non-overlapping areas of each robot, they can be directly transmitted to the global semantic map $\mathcal{M}_t^{(\mathcal{R})}$. For overlapping areas, robots may have different understandings because each robot is fully autonomous at the single robot level. Therefore, an important task of map fusion is to diffuse these local semantic map labels into the global map without prior information.

Define overlapping areas in the local map $m_t^{(r_k)}$ consists of N_{r_k} voxels $C_t^{(r_k)} \triangleq \{v_i^{(r_k)}\}_{1 \le i \le N_{r_k}}$, and define overlapping areas in the global map $\mathcal{M}_t^{(\mathcal{R})}$ consists of $N_{\mathcal{R}}$ voxels $C_t^{(\mathcal{R})} = \bigcup\{C_t^{(r_k)}\} \triangleq \{v_j^{(\mathcal{R})}\}_{1 \le j \le N_{\mathcal{R}}}$. Therefore, the problem is to infer the overlapping areas of the global map from the overlapping areas of the local map:

$$p(\mathcal{M}_t^{(\mathcal{R})} | \{m_t^{(r_k)}\}_{r_k \in \mathcal{R}}) = p(C_t^{(\mathcal{R})} | \{C_t^{(r_k)}\}_{r_k \in \mathcal{R}}) \tag{9}$$

*1) **Graph Model**:* Inspired by [25], we implement the propagation of local map voxels label $\lambda_i^{(r_k)}$ to global map voxels label $\lambda_j^{(\mathcal{R})}$ using label diffusion. The graph model G is formed by nodes that represent both local and global voxels. Only the labels of the global voxels are of interest, so there are two different types of edges in the graph depending on the connected nodes.

To allow the flow of information from local map to global map in our graph, the first type of edge is defined as local-to-global edge. For each local map overlapping areas $C_t^{(r_k)}$, a subgraph $G^{r_k \to \mathcal{R}}$ of size $N_{\mathcal{R}} \times N_{r_k}$ is constructed:

$$G_{ji}^{r_k \to \mathcal{R}} = \begin{cases} \omega \cdot \sum_{l \in \mathcal{L}} \sigma_i^{(r_k)l} & \text{if } \Phi(v_i^{(r_k)}) == v_j^{(\mathcal{R})} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $\Phi(v_i^{(r_k)})$ represents the transformation of voxels $v_i^{(r_k)}$ to the global coordinate according to the robot r_k global localization. ω is a hyperparameter that controls the amount of information flowing from local to global. $\sigma_i^{(r_k)}$ describes the observation quality of robot r_k at the voxel $v_i^{(r_k)}$, representing the confidence that $v_i^{(r_k)}$ is constructed accurately. Therefore, the subgraph $G^{r_k \to \mathcal{R}}$ fuses local semantic maps weighted by credibility, which can improve the confidence level and accuracy of the global map $\mathcal{M}_t^{(\mathcal{R})}$.

The second type of edge is defined as global-to-global edge, which optimizes the label of the global voxel. A subgraph $G^{\mathcal{R} \to \mathcal{R}}$ of size $N_{\mathcal{R}} \times N_{\mathcal{R}}$ is constructed:

$$G_{jj'}^{\mathcal{R} \to \mathcal{R}} = \begin{cases} 1 & \text{if } j = j' \\ k_0(v_j^{(\mathcal{R})}, v_{j'}^{(\mathcal{R})}) & \text{if } ||v_j^{(\mathcal{R})}, v_{j'}^{(\mathcal{R})}|| \le D \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where kernel function k_0 has the same form as (5). In our formulation, Euclidean distance measures the similarity of voxels, which can be used to optimize some mislabeled voxels. In addition, $G^{\mathcal{R} \to \mathcal{R}}$ breaks the independence between voxels again, which is consistent with the single robot level continuous mapping.

Combining these subgraphs, the complete graph G of size $(N_{\mathcal{R}} + \sum_{k=1}^{K} N_{r_k}) \times (N_{\mathcal{R}} + \sum_{k=1}^{K} N_{r_k})$ is defined as:

$$G = \begin{bmatrix} G^{\mathcal{R} \to \mathcal{R}} & G^{r_1 \to \mathcal{R}} & \cdots & G^{r_K \to \mathcal{R}} \\ 0 & I^{r_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I^{r_K} \end{bmatrix} \tag{12}$$

where I^{r_k} is an identity matrix of size $N_{r_k} \times N_{r_k}$.

*2) **Label Diffusion**:* After constructing the graph model G, a label matrix Z is introduced as:

$$Z = [Z^{\mathcal{R}} \quad Z^{r_1} \quad \cdots \quad Z^{r_K}]^T \tag{13}$$

where $Z^{\mathcal{R}}$ of size $N_{\mathcal{R}} \times L$ represents the probabilistic semantic labels of the global voxels and is initialized to a zero

TABLE II

IOU ON SEMANTICKITTI DATASET FOR CONTAINED SEMANTIC CLASSES

Seq.	Method	Car	Road	Sidewalk	Building	Fence	Vegetation	Trunk	Terrain	Pole	**Average**
04	RangeNet++ [26]	91.32	96.87	76.72	49.55	68.89	79.91	12.53	1.31	29.07	56.24
	OctoMap Local Map [3]	90.24	97.28	76.05	55.21	71.03	81.45	13.54	1.69	42.46	58.77
	Yue et al. Global Map [17]	91.34	97.71	76.70	59.45	73.57	82.15	13.83	1.77	45.58	60.23
	Ours Local Map	**93.08**	**98.14**	**78.81**	53.78	71.30	83.18	15.64	1.92	48.60	60.49
	Ours Global Map	83.25	97.94	77.18	**63.39**	**76.55**	**85.18**	**17.85**	**1.96**	**59.95**	**62.58**
05	RangeNet++ [26]	91.33	**95.57**	78.04	86.66	76.10	71.19	22.07	55.58	—	72.07
	OctoMap Local Map [3]	91.91	95.53	79.41	88.71	76.01	73.45	28.93	57.07	—	73.88
	Yue et al. Global Map [17]	93.35	95.56	79.73	89.57	76.47	74.38	31.09	57.73	—	74.74
	Ours Local Map	95.65	95.16	79.95	90.43	78.48	76.50	33.18	61.61	—	76.37
	Ours Global Map	**96.36**	95.29	**80.08**	**91.50**	**78.38**	**76.97**	**35.28**	**62.42**	—	**77.03**

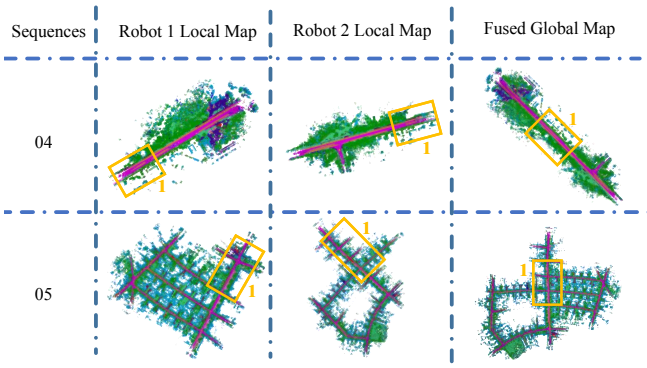

Fig. 4. Collaborative continuous semantic occupancy mapping on SemanticKITTI dataset. The overlapping areas are marked by yellow boxes.

matrix. Z^{r_k} of size $N_{r_k} \times L$ characterizes the probabilistic semantic labels of the local voxels of map r_k.

Label diffusion can be computed by iterative matrix multiplication:

$$Z \leftarrow G \times Z \tag{14}$$

which takes the maximum number of iterations as the termination condition to meet the real-time requirements. As expected, only the global voxel labels matrix $Z^{\mathcal{R}}$ changes:

$$Z^{\mathcal{R}} \leftarrow \left(G^{\mathcal{R}\to\mathcal{R}} \times Z^{\mathcal{R}} + \sum\nolimits_{k=1}^{K} \left(G^{r_k\to\mathcal{R}} \times Z^{r_k} \right) \right) \tag{15}$$

After label diffusion, we extract the fused global voxels label matrix $Z^{\mathcal{R}}$. Each row $Z_j^{\mathcal{R}}$ is identified as semantic count tuple $\sigma_j^{(\mathcal{R})}$. Then the probability label $\lambda_j^{(\mathcal{R})}$ of each global voxel $v_j^{(\mathcal{R})}$ can be deduced as:

$$\lambda_j^{(\mathcal{R})} = \frac{Z_j^{\mathcal{R}}}{\sum_{l' \in L} Z_{jl'}^{\mathcal{R}}}, \quad \lambda_j^{(\mathcal{R})l} = \frac{Z_{jl}^{\mathcal{R}}}{\sum_{l' \in L} Z_{jl'}^{\mathcal{R}}} \tag{16}$$

Finally, the global semantic map $\mathcal{M}_t^{(\mathcal{R})}$ is constructed by fusing all local semantic maps $m_t^{(r_k)}$.

V. EXPERIMENTAL RESULTS

A. SemanticKITTI Dataset

SemanticKITTI [27] is a large-scale outdoor semantic point cloud dataset labeled based on the KITTI odometry dataset. The dataset contains a total of 22 sequences, but only the first 11 sequences provide ground-truth semantic labels, so we randomly selected sequences 04 and 05 for the experiment. We divide each sequence into two segments based on odometry, simulating different observations of two robots. Instead of using ground truth, we take the labels predicted by point cloud segmentation network RangeNet++ [26] as input. The resolution of all maps is set as 0.1 m.

The results constructed by our algorithm are shown in Fig. 4. The generated global semantic map is fused from two local maps. To highlight the difference among different mapping methods, we select 20 scans of point clouds at the overlapping areas of the local map for quantitative analysis. The labels of the map voxels are back-projected to the point cloud as the label prediction results. Tab. II summarizes the IoU of various semantic labels. All maps can improve prior segmentation by fusing multiple scans point clouds. Both our local and global map accuracy is greatly improved due to the introduction of local spatial associations. From another perspective, the global maps have better performance than the local maps because they optimize certain wrong labels through information fusion. The introduction of map credibility and global-to-global edges further optimizes the quality of our global maps.

B. Simulation Environment

To validate the proposed algorithm, we generated several different simulation scenes in the Gazebo platform. The resolution of the map is set to 0.2 m in the experiments. Fig. 5 shows the results for two sets of large-scale environments. As in [14], the uncertainty of voxel v_i is measured by the variance of the posterior distribution:

$$\mathbb{V}(v_i^l) = \frac{\lambda_i^l(1-\lambda_i^l)}{\sum_{l\in\mathcal{L}} \sigma_i^l + 1} \tag{17}$$

The average variance of overlapping areas and mapping efficiency of both local and global maps are listed in Tab.

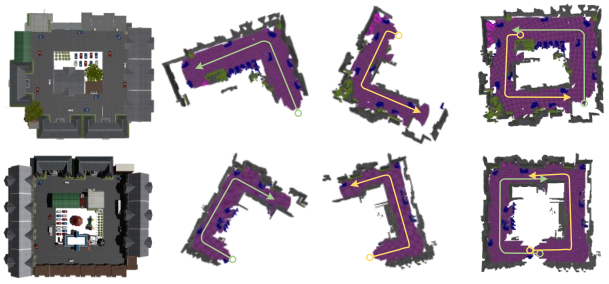

Fig. 5. Collaborative continuous semantic occupancy mapping results in simulation experiments.

TABLE III
VARIANCE AND EFFICIENCY IN SIMULATION EXPERIMENTS

No.	Map	Variance($\times 10^{-4}$)	Efficiency (m^3/s)
1	Robot 1 Local Map	23.93	81.83
	Robot 2 Local Map	28.68	96.69
	Global Map	**8.06**	**173.10**
2	Robot 1 Local Map	23.60	84.01
	Robot 2 Local Map	15.62	88.13
	Global Map	**6.22**	**169.17**

III. The global map has lower variance because it enhances the confidence level by fusing the observations of the two robots, which can improve the safety of robot exploration. Reasonably, since the number of robots $K = 2$, the mapping efficiency of the global map is almost twice that of the local maps. The addition of more robots can greatly improve the efficiency of exploring large-scale scenes.

VI. CONCLUSION

This paper has established a hierarchical and dense collaborative continuous semantic occupancy mapping algorithm (HD-CCSOM). At the single robot level, the local continuous semantic occupancy maps are constructed through multi-entropy kernel inference that introduces local spatial associations. At the collaborative robots level, a label diffusion method is applied to propagate labels from the local maps to the global map. The results have demonstrated that the proposed algorithm achieves high accuracy and efficiency. In summary, this paper provides a new perspective on collaborative semantic perception. In the future, incremental global map fusion with real-time multi-robots operation will be considered.

REFERENCES

[1] R. Shrestha *et al.*, "Learned map prediction for enhanced mobile robot exploration," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 1197–1204. IEEE, 2019.

[2] D. Wang *et al.*, "Behavioral decision-making of mobile robot in unknown environment with the cognitive transfer," *Journal of Intelligent & Robotic Systems*, vol. 103, no. 1, pp. 1–22, 2021.

[3] A. Hornung *et al.*, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous robots*, vol. 34, no. 3, pp. 189–206, 2013.

[4] M. Wen *et al.*, "A robust sidewalk navigation method for mobile robots based on sparse semantic point cloud," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, to be published, 2022.

[5] Y. Yue *et al.*, "Collaborative semantic understanding and mapping framework for autonomous systems," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 2, pp. 978–989, 2020.

[6] Y. Yue *et al.*, "Probabilistic 3d semantic map fusion based on bayesian rule," in *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, pp. 542–547. IEEE, 2019.

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

[8] S. Sengupta *et al.*, "Semantic octree: Unifying recognition, reconstruction and representation via an octree constrained higher order mrf," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1874–1879. IEEE, 2015.

[9] Z. Zhao *et al.*, "Building 3d semantic maps for mobile robots using rgb-d camera," *Intelligent Service Robotics*, vol. 9, no. 4, pp. 297–309, 2016.

[10] B.-s. Kim *et al.*, "3d scene understanding by voxel-crf," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1425–1432, 2013.

[11] S. T. O Callaghan *et al.*, "Gaussian process occupancy maps," *The International Journal of Robotics Research*, vol. 31, no. 1, pp. 42–62, 2012.

[12] F. Ramos *et al.*, "Hilbert maps: Scalable continuous occupancy mapping with stochastic gradient descent," *The International Journal of Robotics Research*, vol. 35, no. 14, pp. 1717–1730, 2016.

[13] T. Shan *et al.*, "Bayesian generalized kernel inference for terrain traversability mapping," in *Conference on Robot Learning*, pp. 829–838. PMLR, 2018.

[14] L. Gan *et al.*, "Bayesian spatial kernel smoothing for scalable dense semantic mapping," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 790–797, 2020.

[15] S. Saeedi *et al.*, "Group mapping: A topological approach to map merging for multiple robots," *IEEE Robotics & Automation Magazine*, vol. 21, no. 2, pp. 60–72, 2014.

[16] Y. Yue *et al.*, "A multilevel fusion system for multirobot 3-d mapping using heterogeneous sensors," *IEEE Systems Journal*, vol. 14, no. 1, pp. 1341–1352, 2019.

[17] Y. Yue *et al.*, "A hierarchical framework for collaborative probabilistic semantic mapping," in *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 9659–9665. IEEE, 2020.

[18] G. Peng *et al.*, "LSDNet: A Lightweight Self-Attentional Distillation Network for Visual Place Recognition," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, to be published, 2022.

[19] G. Peng *et al.*, "Attentional pyramid pooling of salient visual residuals for place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 885–894, Oct. 2021.

[20] S. Jamieson *et al.*, "Multi-robot distributed semantic mapping in unfamiliar environments through online matching of learned representations," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8587–8593. IEEE, 2021.

[21] J. Zhang *et al.*, "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)*, pp. 1039–1044. IEEE, 2018.

[22] A. Melkumyan *et al.*, "A sparse covariance function for exact gaussian process inference in large datasets," in *Twenty-first international joint conference on artificial intelligence*, 2009.

[23] Y. Yue *et al.*, "Cosem: Collaborative semantic map matching framework for autonomous robots," *IEEE Transactions on Industrial Electronics*, 2021.

[24] S. Jin *et al.*, "SectionKey: 3-D Semantic Point Cloud Descriptor for Place Recognition," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, to be published, 2022.

[25] R. Mascaro *et al.*, "Diffuser: Multi-view 2d-to-3d label diffusion for semantic scene segmentation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 589–13 595. IEEE, 2021.

[26] A. Milioto *et al.*, "Rangenet++: Fast and accurate lidar semantic segmentation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4213–4220. IEEE, 2019.

[27] J. Behley *et al.*, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9297–9307, 2019.