S-MKI: Incremental Dense Semantic Occupancy Reconstruction Through Multi-Entropy Kernel Inference

Yinan Deng[1], Meiling Wang[1], Danwei Wang[2], *Fellow, IEEE*, and Yufeng Yue[1*]

***Abstract*— Autonomous robots are often required to acquire high-level prior knowledge by continuously reconstructing the semantics and geometry of the surrounding scene, which is the basis of exploration and planning. Most existing continuous semantic mapping algorithms cannot distinguish potential differences in voxels, resulting in an over-inflated map. Furthermore, fixed-size query ranges introduce high computational complexity. Based on the limitation of over-inflation and inefficiency, this paper proposes a novel incremental continuous semantic occupancy mapping algorithm (S-MKI). The key innovation of this work comes from the two models in the preprocessing stage. On the one hand, Redundant Voxel Filter Model utilizes context entropy to filter out redundant voxels to improve the confidence of the final map, where objects have accurate boundaries with sharp edges. On the other hand, Adaptive Kernel Length Model adaptively adjusts the kernel length with class entropy, which reduces the inherent amount of training data. The final multi-entropy kernel inference function is formulated to integrate these two models to infer sparse noisy sensor data into dense accurate 3D maps. Experimental results conducted in both indoors and outdoors datasets validate that S-MKI outperforms existing methods.**

I. Introduction

The essence of robot mapping is to employ sparse noisy sensor observations to construct a dense accurate representation, which is regarded as a fundamental problem in robotics [1]. Currently, the most widely used mapping technique is occupancy grid map [2]. Grid mapping approaches assume that the voxels are statistically independent, and therefore generate discrete maps, which contradict the fact that the surface of the object is continuous in the real world. Recent success in Bayesian kernel inference has boosted the development of continuous mapping. They incorporate local spatial correlations into the mapping model, which can infer the continuous surface from sparse sensor data. However, they neglect the potential differences between voxels and treat all voxels with the same importance, so the voxels next to the object are misclassified to be occupied, which leads to over-inflation of objects. Such an over-inflated map is difficult to apply to robot navigation tasks, because traversable free space might be falsely blocked [3], [4]. In addition, as robots are required to perform more intelligent tasks, incorporating semantic information can further help them distinguish object categories and allow a higher level of environmental representation [5], [6]. Therefore, the main objective of this paper is to design a novel continuous semantic occupancy mapping algorithm that can mitigate over-inflation while improving efficiency.

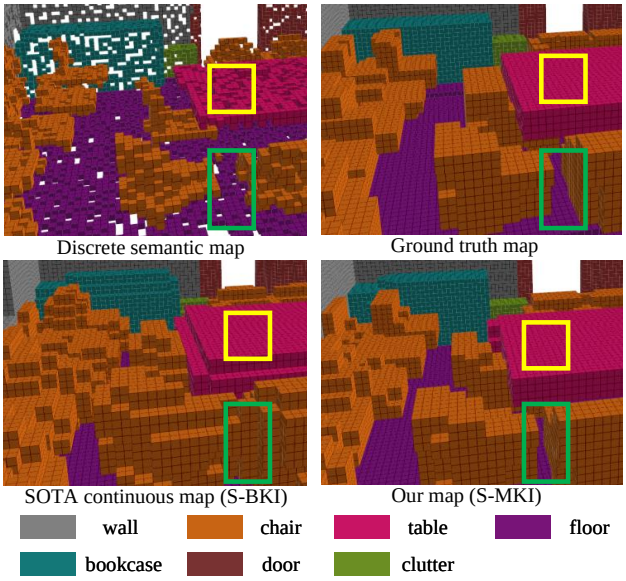


Fig. 1. The demonstration of a conference room of Stanford 2D-3D Semantic Dataset. Our map restores a smooth surface for the objects while generating precise boundaries with sharp edges.

The first challenge is to infer the voxels that are worth filling, so as to mitigate the over-inflation phenomenon of existing continuous mapping methods. As shown in Fig. 1, the discrete map only recovers the area hit by the sensor observations, leaving many loopholes. These mapped voxels are defined as *observed voxels* in this paper. Other unknown voxels in the map can be divided into two types: the voxels in the loopholes due to the lack of observation data are called *inactive voxels* (Fig. 1 yellow box), while those in the free space outside the objects are called *redundant voxels* (Fig. 1 green box). Continuous mapping is expected to only fill in *inactive voxels*, generating a representation similar to the ground truth map. However, SOTA continuous mapping method S-BKI [7] did not distinguish unknown voxels, building an over-inflated map by falsely filling in *redundant voxels* (Fig. 1). In order to accurately reconstruct the scene, the Redundant Voxel Filter Model is proposed to filter out redundant voxels with context entropy, which aims to increase the confidence level in the inference process.

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

[1]Yinan Deng, Meiling Wang and Yufeng Yue are with School of Automation, Beijing Institute of Technology, Beijing, 100081, China.

[2]Danwei Wang is with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

The second challenge is to reduce the computational complexity of continuous semantic occupancy mapping, thereby improving efficiency. Current continuous approaches adopt fixed kernel length, which is n times of the voxel size. This operation will increase the computational complexity by n^3 compared to discrete approaches. To reduce the time cost, the Adaptive Kernel Length Model is proposed to adjust the kernel length adaptively by introducing class entropy, which is the measurement of the overall uncertainty of a voxel.

In summary, over-inflation and inefficiency are two challenging problems of continuous semantic occupancy mapping. In this paper, a novel incremental continuous semantic occupancy mapping algorithm (S-MKI) is proposed. We mathematically formulate the overall continuous semantic occupancy mapping problem and derive its probabilistic models.

II. Related Works

A. Semantic Mapping

With the rapid development of deep learning, semantic mapping has attracted increasing attention [8]. Early semantic mapping methods [9] directly use semantic images to perform mapping. The authors in [10] use the Bayesian framework to filter probabilistic segmentation from multiple views in a voxel-based 3D map. In [11], the street-level image label estimates are aggregated to annotate the 3D volume. These methods are the pioneers of semantic mapping, but they lack further optimization. To optimize incorrect voxel labels, CRF has become a research hotspot [12], which can simulate the long-distance relationships in a region, such as 2D superpixels [13] or 3D supervoxels [14]. In [15], a novel high-order CRF model is applied to optimize 3D grid labels.

Various methods mentioned above promoted the development of semantic mapping. However, these methods assume that the voxels are independent and do not reconstruct the continuous surface of the objects.

B. Continuous Mapping

To construct a smoother occupancy map, many methods have attempted to relax the assumption that the voxels are independent, such as GPmap [16], Hilbert map [17], etc. GPmap [16] introduces a dependency relationship between points as the non-parametric Bayesian inference process. However, the $\mathcal{O}(n^3)$ computational complexity has limited its application to large-scale online mapping [18]. Hilbert map [17] makes use of fast kernel approximations to enable faster training in $\mathcal{O}(n)$ time. Recently, Bayesian kernel inference with $\mathcal{O}(\log n)$ computational complexity has begun to gain attention. BGKOctoMap [19] innovatively applies the sparse kernel and Bayesian non-parametric inference to improve efficiency. More recently, S-BKI [7] extends [19] to 3D semantic mapping, which enriches the map information.

Unfortunately, the above methods ignore the potential difference of voxels, which results in over-inflated maps. Although AKImap [20] uses a bandwidth matrix to alleviate this problem, the additional computational effort is prohibitive. These have become the reasons that limit the wide application of continuous mapping.

III. Incremental Continuous Semantic Occupancy Mapping

A. Algorithm Framework and Problem Definition

The framework of the S-MKI algorithm is depicted in Fig. 2, which consists of three main modules. In the Redundant Voxel Filter Model (RVFM), redundant voxels are filtered out, leaving inactive and observed voxels for continuous inference. In the Adaptive Kernel Length Model (AKLM), class entropy composed of two sub-entropies is introduced to adjust the kernel length, which determines the range of local spatial associations. Finally, the estimation from sensor observations to a continuous semantic map is achieved by Multi-entropy Kernel Inference, which incorporates the information conveyed by RVFM and AKLM.

Considering a robot operating in a completely unknown environment and attempting to reconstruct the surroundings, the problem can be defined as follows:

Problem Definition: *Given a robot r with camera observations $I_{1:t}$, 3D LiDAR observations $L_{1:t}$ and robot trajectory $O_{1:t}$, the objective is to estimate the continuous semantic occupancy map $\mathcal{M}_t$.*

$$p(\mathcal{M}_t|I_{1:t}, L_{1:t}, O_{1:t}) \tag{1}$$

The solution of the problem corresponds to the Maximum A Posterior (MAP) estimation of (1). For input, we have $I_t \in \mathbb{R}^2$ in 2D, $L_t \in \mathbb{R}^3$ in 3D and $O_t \in SE(3)$ in 3D. For output, the dense semantic map $\mathcal{M} \triangleq \{v_j\}_{j=1}^{N_\mathcal{M}}$ consists of a set of voxels. Each voxel v_j contains the 3D coordinate of the center (v_j^x, v_j^y, v_j^z), associated with a tuple $\lambda_j = (\lambda_j^1, \lambda_j^2, \cdots \lambda_j^K)$ to store probabilistic semantic labels, where K is the total number of semantic classes and $\sum_{k=1}^{K} \lambda_j^k = 1$. In this paper, *free* is modeled as a special semantic class to represent the traversable space.

At time t, the RGB image I_t is fed into segmentation network [21]. For each pixel, the output is a one-hot encoded measurement tuple $c_i = (c_i^1, c_i^2, \cdots c_i^K)$. The semantic labels can be transmitted from pixels to LiDAR points by projection [22], where the parameters are calibrated by [23]. Therefore, (1) can be rewritten as:

$$p(\mathcal{M}_t|I_{1:t}, L_{1:t}, O_{1:t}) = p(\mathcal{M}_t|L_{s_{1:t}}) \tag{2}$$

The semantic point cloud $L_s \triangleq \{p_i\}_{i=1}^{N_{L_s}}$ consists of a series of semantic points p_i referred by coordinates (p_i^x, p_i^y, p_i^z), which are associated with semantic label c_i. Alternatively, the problem can be refined as:

Given semantic points and labels $\{p_i, c_i\}_{i=1}^{N_{L_s}}$, the objective is to estimate probabilistic semantic labels λ_j of each voxel v_j.

$$p(\mathcal{M}_t|L_{s_{1:t}}) = \prod_{j=1}^{N_\mathcal{M}} \prod_{i=1}^{N_{L_s}} p(\lambda_j|\, v_j, p_i, c_i) \tag{3}$$

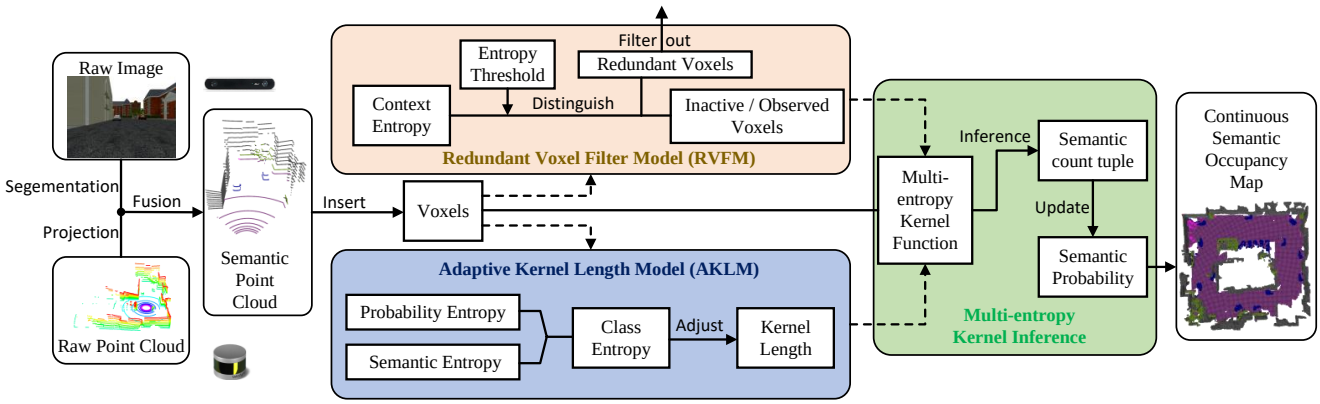
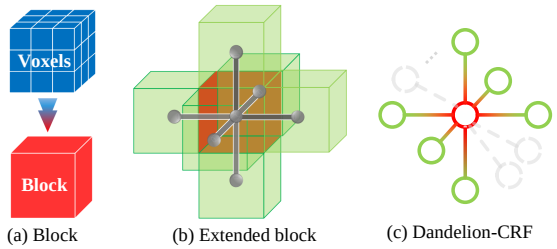


Fig. 2. The framework of incremental continuous semantic occupancy mapping algorithm S-MKI.



(a) Block (b) Extended block (c) Dandelion-CRF

Fig. 3. The illustration of voxels and blocks. (a) Red is a block b_J composed of several blue voxels $\{v_j\}$. (b) The green blocks $\{b_x\}$ and current block b_J together form the extended blocks B_J. (c) Dandelion-CRF extracted from the extended blocks B_J. The gray part implies its flexible scalability.

B. Redundant Voxel Filter Model

As stated before, previous continuous mapping methods cannot clearly distinguish voxels, resulting in overfitting of the final continuous map. The Redundant Voxel Filter Model is designed to address this problem.

As shown in Fig. 3a, in order to improve semantic accuracy and inference efficiency, *block* is introduced as an intermediate layer between *map* and *voxel*. Each block $b_J \triangleq \{v_j\}_{j\in J}$ is a small semantic octree, comprised of several adjacent voxels v_j. In consideration of the time cost, the filtering of redundant voxels is performed in block units, so all the voxels v_j in the block b_J will inherit the attributes of the parent block. To clarify the state around the current block b_J, we also introduce extended blocks $B_J \triangleq \{b_J, \{b_x\}\}$ (Fig. 3b), which consists of the current block b_J and some blocks $\{b_x\}$ around b_J. Each block b_J can independently construct its own extension block B_J.

We condense the extended blocks into a graph model (Fig. 3c), with blocks in the extended blocks as nodes, and the connection between the current block b_J and surrounding blocks $\{b_x\}$ as edges. This graph is called Dandelion-CRF because of its highly recognizable and flexible scalability structure. Given the observation D, the context entropy $\mathbf{E}_{con}$ is described as the conditional probability of the central node b_J:

$$\mathbf{E}^J_{con} = P(b_J \sim 1|D) \tag{4}$$

where, $b_J \sim 1$ indicates that b_J should be filled to enhance the consistency of the map. It is worth noting that filling is to use the spatial association to populate current observation.

There are two kinds of cliques in Dandelion-CRF: One is a single node $\{b_k\}$ and the other is a pair of adjacent nodes $\{b_k, b_l\}$, where k and l are index variables. By selecting the exponential potential function and introducing the feature function, the conditional probability is defined as:

$$P(b_J \sim 1|D) = \frac{1}{Z(D)} \exp(E(b_J \sim 1|D)) \tag{5}$$

$$E(b_J \sim 1|D) = \psi(b_J) + \sum_x (\psi(b_x)\psi(b_J, b_x)) \tag{6}$$

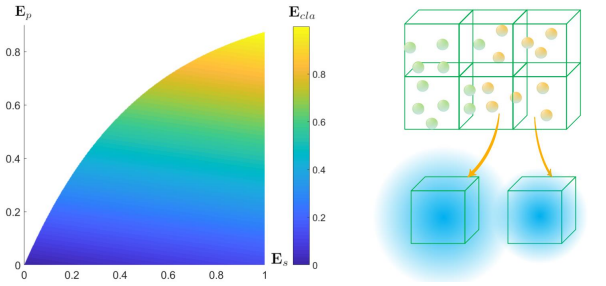
where $Z(D)$ is the partial function for normalization, the status feature function $\psi(b_k)$ and the transition feature function $\psi(b_k, b_l)$ describe the influence of the observation sequence and adjacent nodes, respectively. In our formulation, $\psi(b_k)$ obtains different values according to whether the block is observed or not. $\psi(b_k, b_l)$ takes the Radial Basis Function (RBF) of the Euclidean distance between blocks. In (7) and (8), ω_1 and ω_2 are hyperparameters to control the amount of information transmitted, and s is the resolution of the block.

$$\psi(b_k) = \begin{cases} \omega_1 & \exists p_i \in b_k \\ 0 & \text{Otherwise} \end{cases} \tag{7}$$

$$\psi(b_k, b_l) = \frac{\omega_2}{\omega_1} \exp(-\frac{\|b_k - b_l\|^2}{2s^2}) \tag{8}$$

$\mathbf{E}_{con}$ reveals potential differences between voxels that can be used as indicators of differentiation. Taking $\mathbf{T}_{con}$ as the entropy threshold, the voxels contained in the block with context entropy less than $\mathbf{T}_{con}$ are redundant, often located in the gap between two objects or outside the edge of the object. RVFM will filter out these redundant voxels, while preserving observed and inactive voxels to estimate a more accurate map. This operation is realized by the filtering factor f_J transmitted to the multi-entropy kernel inference module.

$$f_J = [\mathbf{E}^J_{con} \geq \mathbf{T}_{con}] \tag{9}$$



(a) Class entropy (b) Adaptive kernel length

Fig. 4. (a) The value of class entropy $\mathbf{E}_{cla}$ in the sub-entropies plane. (b) Adaptive kernel length. The voxels with inconsistent observations are assigned a larger kernel length to increase their confidence level by querying surrounding measurements.

C. Adaptive Kernel Length Model

The kernel length is the key to mapping efficiency, because it determines the query range. The Adaptive Kernel Length Model is designed to assign appropriate kernel lengths to voxels. We continue to use *block* as the unit, and voxels in the same block are assigned the same kernel length.

Class entropy $\mathbf{E}_{cla}$ is introduced to measure overall uncertainty of the voxel. It contains two sub-entropies: one is the probability entropy $\mathbf{E}_p$, and the other is semantic entropy $\mathbf{E}_s$. On the one hand, probability entropy $\mathbf{E}_p$ reflects the proportion of number, which is defined as:

$$\mathbf{E}_p = \frac{\mathfrak{n}_{all} - \mathfrak{n}_{max}}{\mathfrak{n}_{all}} \tag{10}$$

where $\mathfrak{n}_{max}$ is the number of semantic points that account for the largest number among all semantic classes, and $\mathfrak{n}_{all}$ is the total number of semantic points. On the other hand, semantic entropy $\mathbf{E}_s$ describes the diversity of semantic labels. Defining $\mathfrak{k}$ to indicate the number of semantic labels contained in block b_J, semantic entropy $\mathbf{E}_s$ is defined as:

$$\mathbf{E}_s = \log_K(\mathfrak{k}) = \frac{\ln(\mathfrak{k})}{\ln(K)} \tag{11}$$

Class entropy $\mathbf{E}_{cla}$ is defined in (12) to combine two sub-entropies. Probability entropy $\mathbf{E}_p$ is dominant because it integrates part information of semantic entropy $\mathbf{E}_s$ (See Fig. 4a). When there is no observation in the block, it will have the largest class entropy. This situation also occurs when points with any semantic labels fall evenly into the block.

$$\mathbf{E}_{cla}^J = \begin{cases} \dfrac{\mathfrak{n}_{all} - \mathfrak{n}_{max}}{\mathfrak{n}_{all}} + \dfrac{\log_K(\mathfrak{k})}{K} & \exists p_i \in b_J \\ 1 & otherwise \end{cases} \tag{12}$$

As shown in Fig. 4b, larger class entropy means higher uncertainty, requiring a larger query range to ensure map accuracy. Therefore, for voxel v_j in block b_J, the kernel length with bounds L_{min} and L_{max} is adjusted to:

$$L_J = L_{min} + \mathbf{E}_{cla}^J (L_{max} - L_{min}) \tag{13}$$

D. Multi-entropy Kernel Inference

The efficacy of the RVFM and AKLM needs to be exerted through multi-entropy kernel inference, which essentially converts sensor observations into updated maps. Different from the classical voxel probability update model, the multi-entropy kernel inference model is derived based on the counting sensor model. According to the Bayesian rule, (3) can be decomposed into:

$$p(\lambda_j |\, v_j, p_i, c_i) \propto p(\lambda_j) p(c_i |\, v_j, p_i, \lambda_j) \tag{14}$$

For incremental Bayesian inference, likelihood probability is modeled as a Categorical distribution $Cat(\lambda_j^1, \lambda_j^2, \cdots \lambda_j^K)$. And both prior probability and posterior probability satisfy Dirichlet distribution $Dir(K, \sigma_0)$ and $Dir(K, \sigma_j)$, where $\sigma_0 = \{\sigma_0^1, \sigma_0^2, \cdots \sigma_0^k\}$ and $\sigma_j = \{\sigma_j^1, \sigma_j^2, \cdots \sigma_j^k\}$ are the distribution parameters.

In order to break the independence of voxels, the kernel function k operating on 3D spatial $\mathcal{X} \times \mathcal{X} \to [0, 1]$ is utilized to create an extended likelihood. After simplification, the relationship between the two Dirichlet distribution parameters σ_0 and σ_j can be obtained:

$$\sigma_j = \sigma_0 + k(v_j, p_i) c_i \tag{15}$$

where σ_j is the weighted count of the semantic points, which is known as the semantic count tuple of voxel v_j. The (15) states that σ_j counts not only the semantic points that fall into the current voxel v_j, but also the adjacent semantic points with the kernel function as the weight.

In order to reduce the computational complexity, we chose the sparse kernel function $k_0(v_j, p_i)$ [24] as a template, which is recommended in previous works:

$$k_0(v_j, p_i) = \mathop{\mathbf{I}}_{d<L} \varepsilon_0 \left[\frac{(2 + \cos(2\pi \frac{d}{L})(1 - \frac{d}{L}))}{3} + \frac{\sin(2\pi \frac{d}{L})}{2\pi} \right] \tag{16}$$

where $\mathbf{I}$ represents the indicator function, $d = \|v_j - p_i\|$, L is the kernel length, and ε_0 is the scale factor.

Incorporating the proposed RVFM (9) and AKLM (13) into the (16), the multi-entropy kernel function $k_e(v_j, p_i)$ is derived as:

$$k_e(v_j, p_i) = f_J k_0(v_j, p_i)_{L \to L_j} \tag{17}$$

Inserting semantic point clouds $L_{s_{1:t}}$, the probabilistic semantic label λ_j of the voxel v_j is the closed-form expected value of the posterior Dirichlet:

$$\lambda_j^k = \frac{\sigma_j^k}{\sum_{m=1}^{K} \sigma_j^m} = \frac{\sigma_0^k + \sum_{i=1}^{N_{L_s}} k_e(v_j, p_i) c_i^k}{\sum_{m=1}^{K} \left(\sigma_0^m + \sum_{i=1}^{N_{L_s}} k_e(v_j, p_i) c_i^m \right)} \tag{18}$$

In summary, the mapping problem defined in (3) has been transformed into probabilistic solution as (18). When new sensor observations are obtained, the map can be updated by calculating (18) incrementally.

Fig. 5. The semantic mapping results on a lounge of Stanford 2D-3D Semantic Dataset.

TABLE I
QUANTITATIVE RESULTS ON A LOUNGE OF STANFORD 2D-3D SEMANTIC DATASET

Scenes	Method	Wall	Chair	Table	Floor	Bookcase	Door	Beam	Clutter	Ceiling	W.-average	Accuracy
Lounge	S-CSM	**36.23**	20.82	15.85	33.76	12.68	—	—	21.04	38.43	27.36	80.32
	S-BKI	24.26	51.97	47.14	25.04	**56.32**	—	—	47.78	28.64	37.49	54.95
	S-MKI	35.98	**62.01**	**49.77**	**62.71**	43.81	—	—	**60.07**	**78.59**	**54.24**	**84.78**

IV. EXPERIMENTAL RESULTS

Comparison Baseline: Due to the limited available continuous semantic mapping algorithms, S-CSM and S-BKI [7] are set as the baselines. For a fair comparison, we rerun these algorithms under the same configuration.

Evaluation Metric: Accuracy is measured by voxel-IoU, w-average and accurateness. Voxel-IoU extends the pixel-IoU from 2D to 3D, which is defined as TP/(TP+FP+FN). W-average is the weighted average of voxel-IoU. Accurateness is defined as the proportion of correctly classified voxels. Efficiency is measured in seconds.

A. Stanford Indoor Dataset

Stanford 2D-3D Semantics Dataset is a large indoor spatial dataset. A lounge in area 3 is selected as a test scene to evaluate the accuracy. The map resolution is set as 0.05 m for all algorithms.

The mapping results are shown in Fig. 5. As can be seen, S-CSM generates a discrete semantic map by only predicting observed voxels. S-BKI confuses redundant voxels and inactive voxels, and the object becomes very thick. The enlarged picture shows the entire chair has been over-inflated and connected to the floor. In contrast, our proposed S-MKI successfully filters out redundant voxels while filling in the inactive voxels, which builds a semantic map that visually has the most similar features to the ground truth.

The quantitative evaluation results of the mapping accuracy are summarized in Tab. I. S-MKI has achieved significant advantages, which is consistent with the visual results. S-BKI has a higher IoU than S-CSM, but has the lowest accuracy due to the overfilling of a large number of redundant voxels. In this experiments, the mapping frequency of S-MKI almost doubled compared to the baseline.

B. SemanticKITTI Outdoor Dataset

SemanticKITTI Dataset is a large outdoor semantic point cloud dataset based on the KITTI odometry dataset, collected from the real world. Sequence 04 is randomly selected for experiments. The map resolution is set as 0.3 m.

The comparison of the mapping results is shown in Fig. 6. The enlarged pictures present part of the ground. Due to inaccurate network segmentation, all the generated maps have some random noises. There are many loopholes and messy semantic labels in the S-CSM map. The reason is that S-CSM does not consider the spatial correlation. S-BKI can remove some noises and fill in the loopholes by smoothing, but it is still not comparable with ours. S-MKI almost removes all noises by applying RVFM and AKLM. More specifically, S-MKI does not overly consider surrounding observations for areas with accurate observations, as this may diffuse surrounding noise.

The quantitative results are summarized in Tab. II. It is obvious that S-MKI has the best performance. Moreover, the accuracy of S-BKI has been greatly improved, and even surpasses S-CSM. The reason is that continuous mapping is more suitable for cluttered outdoor scenes, which have many unknown or fuzzy objects. In outdoor scenes experiments, the time-consuming of S-MKI is only one-fourth of the baseline thanks to the shrinkable kernel length.

V. CONCLUSION

This paper has established a novel incremental continuous semantic occupancy mapping algorithm (S-MKI). More specifically, the proposed Redundant Voxel Filter Model (RVFM) filters out redundant voxels, therefore the representation of objects has accurate boundaries with sharp edges. In addition, the proposed Adaptive Kernel Length Model (AKLM) adjusts kernel length adaptively, which greatly reduces the computational complexity. The multi-entropy

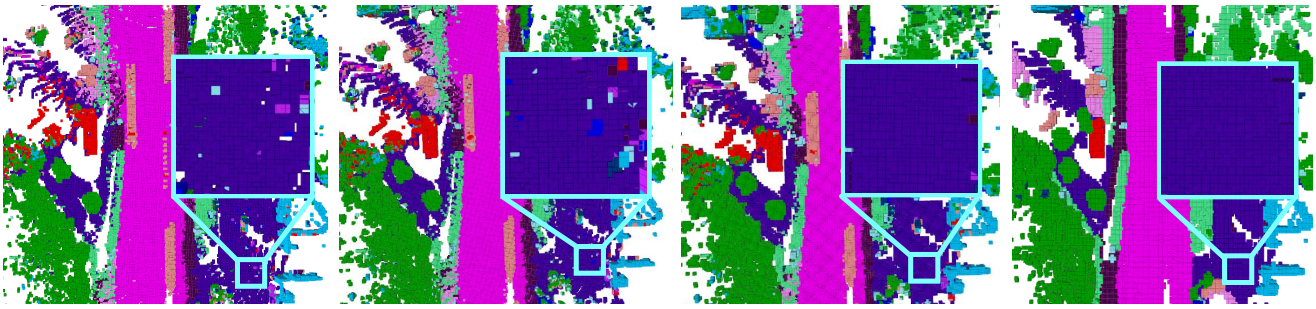


Fig. 6. The semantic mapping results on Sequence 04 of SemanticKITTI Dataset.

TABLE II
QUANTITATIVE RESULTS ON SEQUENCE 04 OF SEMANTICKITTI DATASET

Seq.	Method	Car	Other-vehicle	Person	Road	Parking	Sidewalk	Other-ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic-sign	W.-average	Accuracy
04	S-CSM	7.97	18.67	6.82	25.59	7.13	6.41	20.98	14.23	9.70	17.49	5.32	12.78	8.59	9.86	17.29	76.20
	S-BKI	10.42	22.16	9.03	46.79	11.32	15.30	39.37	23.96	13.88	32.96	6.06	25.15	12.76	14.02	31.96	80.20
	S-MKI	**16.91**	**26.60**	**9.17**	**79.79**	**21.60**	**26.39**	**60.92**	**42.77**	**23.84**	**64.67**	**8.85**	**43.22**	**23.42**	**23.67**	**57.36**	**89.92**

kernel function integrates these two models. The results have demonstrated that the proposed algorithm achieves high accuracy and efficiency. In summary, S-MKI addresses two significant problems in continuous semantic occupancy mapping and provides a new perspective on environmental reconstruction.

REFERENCES

[1] Y. Yue *et al.*, *Collaborative Perception, Localization and Mapping for Autonomous Systems*, vol. 2. Springer Nature, 2020.

[2] 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.

[3] C. Wang *et al.*, "Efficient object search with belief road map using mobile robot," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3081–3088, 2018.

[4] C. Wang *et al.*, "Autonomous robotic exploration by incremental road map construction," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 4, pp. 1720–1731, 2019.

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

[6] 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.

[7] 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.

[8] S. Garg *et al.*, "Semantics for robotic mapping, perception and interaction: A survey," *arXiv preprint arXiv:2101.00443*, 2021.

[9] H. He *et al.*, "Nonparametric semantic segmentation for 3d street scenes," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3697–3703. IEEE, 2013.

[10] J. Stückler *et al.*, "Semantic mapping using object-class segmentation of rgb-d images," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3005–3010. IEEE, 2012.

[11] S. Sengupta *et al.*, "Urban 3d semantic modelling using stereo vision," in *2013 IEEE International Conference on robotics and Automation*, pp. 580–585. IEEE, 2013.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] J. Wang *et al.*, "Fast, accurate gaussian process occupancy maps via test-data octrees and nested bayesian fusion," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1003–1010. IEEE, 2016.

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

[20] Y. Kwon *et al.*, "Adaptive kernel inference for dense and sharp occupancy grids," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4712–4719. IEEE, 2020.

[21] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[22] 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.

[23] C. Guindel *et al.*, "Automatic extrinsic calibration for lidar-stereo vehicle sensor setups," in *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pp. 1–6. IEEE, 2017.

[24] 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.