Bayesian Spatial Kernel Smoothing for Scalable Dense Semantic Mapping

Lu Gan, Ray Zhang, Jessy W. Grizzle, Ryan M. Eustice, and Maani Ghaffari

***Abstract*—This paper develops a Bayesian continuous 3D semantic occupancy map from noisy point clouds by generalizing the Bayesian kernel inference model for building occupancy maps, a binary problem, to semantic maps, a multi-class problem. The proposed method provides a unified probabilistic model for both occupancy and semantic probabilities and nicely reverts to the original occupancy mapping framework when only one occupied class exists in obtained measurements. The Bayesian spatial kernel inference relaxes the independent grid assumption and brings smoothness and continuity to the map inference, enabling to exploit local correlations present in the environment and increasing the performance. The accompanying software uses multi-threading and vectorization, and runs at about 2 Hz on a laptop CPU. Evaluations using multiple sequences of stereo camera and LiDAR datasets show that the proposed method consistently outperforms current baselines. We also present a qualitative evaluation using data collected with a bipedal robot platform on the University of Michigan - North Campus.**

***Index Terms*—Mapping, semantic scene understanding, range sensing, RGB-D perception.**

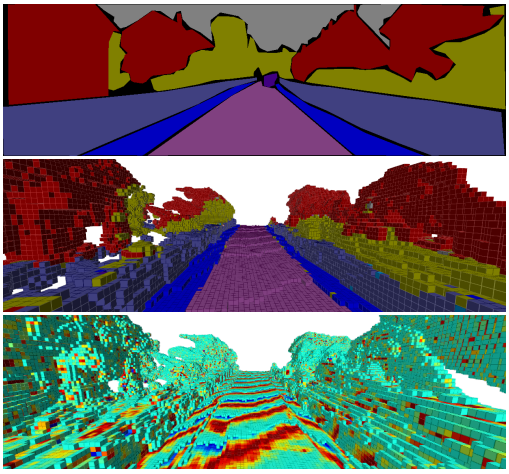

Fig. 1: Qualitative results on KITTI odometry sequence 05 dataset [13]. From top to bottom the figures show the 2D ground truth image, 3D semantic map, and variance map.

I. Introduction

ROBOTIC mapping is the problem of inferring a representation of the robot's surroundings using noisy measurements as it navigates through an environment. This problem is traditionally solved using occupancy grid mapping techniques [1]–[3]. As robotic systems move toward more challenging behaviors in more complex scenarios, such systems require richer maps so that the robot understands the significance of the scene and objects within. Hence, the integration of semantic knowledge into the map has been the focus of robotic research in recent years [4]–[9].

A semantic occupancy map as shown in Fig. 1, besides possessing properties similar to an occupancy grid map, maintains for each cell a set of probabilities of semantic classes. These probabilities are often updated using a Bayes filter [9], [10], and then Conditional Random Fields (CRF) or Markov Random Fields (MRF) are subsequently applied to mitigate discontinuities and inconsistencies in the semantic map [7]–[9], [11], [12]. In principle, CRF models encourage label consistency among neighboring grids in super-voxels [8] or 2D superpixels [9], [12]. However, CRF optimization is only applied as a post-processing step, and therefore, it is unable to predict semantics of partially observed regions in the map.

This work was partially supported by the Toyota Research Institute (TRI), partly under award number N021515. Funding for J. Grizzle was in part provided by TRI and in part by NSF Award No. 1808051.

The authors are with the University of Michigan, Ann Arbor, MI 48109, USA. ganlu@umich.edu, rzh@umich.edu, grizzle@umich.edu, eustice@umich.edu, maanigj@umich.edu

Occupancy grid maps assume the grids are statistically independent. However, a series of investigations on continuous occupancy mapping shows that taking local spatial correlations into account increases mapping performance [14]–[21]. Building on a similar idea, continuous semantic maps [22], [23] can deal with sparse sensor measurements by inferring semantics of partially observed regions from neighboring measurements. Recent work on Bayesian generalized kernel inference for occupancy map prediction (BGKOctoMap) proposed in [21] uses a kernel inference approach to generalize the counting sensor model [24] to continuous maps while maintaining the scalability of the method.

In this paper, we extend BGKOctoMap [21] to continuous semantic mapping where the inference reverts to the original framework when only one occupied class exists. In particular, the contributions of this work are 1) we develop a continuous statistical model for semantic occupancy mapping which models occupancy and semantic probabilities in a unified framework and queries can be made at any resolution; 2) we provide an open-source implementation of the proposed method. The current implementation exploits multi-threading and vectorization and can be run at about 2 Hz using a laptop CPU; 3) we present extensive experiments using both stereo camera and LiDAR data. The evaluations show that the proposed method consistently outperforms state-of-the-art systems.

Related work is given in Section II. Section III presents preliminaries and semantic counting sensor model. Section IV

describes an extension to continuous mapping. Experimental results are presented in Section V. Limitations of this work are discussed in Section VI and Section VII concludes the paper.

II. Related Work

Discrete 3D Semantic Mapping. Early semantic mapping work uses traditional pixel-wise image segmentation methods and directly transfers image labels from 2D to 3D. Labels from multiple images are fused in 3D without any further 3D optimization [10], [25], [26]. He et al. [25] build a semantic octomap by using an MRF for image segmentation and selecting the most frequent label of the 3D points inside each grid. Sengupta et al. [26] build a semantic volumetric map by adopting a CRF for 2D semantic segmentation and assigning labels by a voting scheme. Stückler et al. [10] use random decision forests to segment object classes in images and fuse soft labels in a voxel-based 3D map using a Bayesian update. While these methods are similar to our semantic counting sensor model in a way that the maximum of semantic labels in a 3D element is picked in label fusion, the latter is a closed-form Bayesian inference which outputs the mean and variance of the posterior.

To deal with noisy 2D predictions, 3D CRF optimization has been introduced as a refinement technique and it is widely applied in 3D semantic mapping [7], [11], [27]. In [8], [12], [13], a higher-order dense CRF model is used to further optimize the semantic predictions for 3D elements. Basic CRF models encourage label consistency for adjacent 3D elements, while higher-order dense CRFs can model long-range relationships within a region, such as grids in super-voxels [8] or grids corresponding to 2D superpixels [12], and further improve the mapping performance. More recent work uses deep Convolutional Neural Networks (CNNs) for 2D image segmentation, and follows the same framework for building 3D semantic maps [9], [28]. However, CRF optimization post-processes the inferred occupied grids, which does not change the principle of discrete semantic map inference. In [29], a semantic Simultaneous Localization and Mapping (SLAM) system, SuMa++, builds a surfel-based semantic map using SemanticKITTI dataset [30] as its byproduct. However, surfel-based maps do not model occupied or free space, thus are not used for robot navigation.

Continuous Mapping. Gaussian Process Occupancy Map (GPOM) [14] takes into account the correlation between map points and treats the map inference as a binary classification at an arbitrary resolution. Hilbert maps [18] are more scalable and can be updated in linear time where a logistic regression classifier is trained online through stochastic gradient descent. GPOM has been extended from binary to multi-class case in [22]. However, the complexity of the model grows with the number of data points, resulting in $\mathcal{O}(n^3)$ cost without approximation. The cost also grows with the number of semantic classes as a one-vs.-rest approach is used to build the multi-class classifier. Similarly, Hilbert maps can also be extended to the multi-class maps using a multinomial model. However, as discussed in [21], the logistic regression classifier used by Hilbert map-based approaches does not provide associated uncertainties in probability estimates.

Bayesian Kernel Inference. Bayesian Kernel Inference (BKI) was introduced in [31] as an approximation to Gaussian processes that requires only $\mathcal{O}(\log N)$ computations instead of $\mathcal{O}(N^3)$, where N is the number of training points. It generalizes local kernel estimation to the context of Bayesian inference for the exponential family of distributions. Instead of approximating inference on the model, the approximation is made at the stage of model selection. Assuming latent training parameters are conditionally independent given the target parameters, exact inference on this model is possible for any likelihood function from the exponential family. In [32], BKI is successfully applied to a visual odometry problem for modeling sensor uncertainty. In [33], BKI has been used on a Bernoulli-distributed random event with Beta-distributed prior to model collision in safe high-speed navigation problems and could achieve safe behavior in a novel environment with no relevant training data. BKI was first used in the context of mapping problems in [20], [21], to generalize the discrete counting sensor model [24] to continuous occupancy mapping. Later, the applications of BKI in elevation regression and traversability classification are explored in [34]. Following the same idea, we apply BKI in our semantic counting sensor model and generalize it to continuous semantic mapping. In particular, we use BKI on a Categorical likelihood with a Dirichlet distribution as its conjugate prior.

III. Preliminaries and Semantic Counting Sensor Model

The counting sensor model describes occupancy probability via a Bernoulli likelihood function. It counts for each grid how often a beam has ended in that grid and how often a beam has passed through it. This model has comparable performance to Bayesian updates in occupancy grid mapping [35]. The semantic counting sensor model is its natural generalization from occupancy (binary) mapping to semantic (multi-class) mapping.

Let $\mathcal{K} = \{1, 2, ..., K\}$ be the set of semantic class labels, i.e., K categories, and $\mathcal{X} \subset \mathbb{R}^3$ be the map spatial support. For any map point $x_i \in \mathcal{X}$, we have a one-hot-encoded measurement tuple $y_i = (y_i^1, ..., y_i^K)$, where $y_i^k \geq 0$ and $\sum_{k=1}^{K} y_i^k = 1$. In practice, y_i is the output of a max function computed using the output of a deep network for multi-class classification. The training set (data) can be defined as $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^{N}$.

Assuming map cells are indexed by $j \in \mathbb{Z}^+$, the jth map cell can take on one of K possible categories with the probability of each category separately specified as $\theta_j = (\theta_j^1, ..., \theta_j^K)$, where $\sum_{k=1}^{K} \theta_j^k = 1$. The jth map cell with semantic probability θ_j is described by a Categorical distribution as:

$$p(y_i|\theta_j) = \prod_{k=1}^{K} \left(\theta_j^k\right)^{y_i^k}. \tag{1}$$

In semantic mapping, we seek the posterior over θ_j; $p(\theta_j|\mathcal{D})$.

For incremental Bayesian inference, we adopt a Dirichlet prior distribution over θ_j, given by $Dir(K, \alpha_0)$, as the conjugate prior of the Categorical likelihood, where $\alpha_0 = (\alpha_0^1, ..., \alpha_0^K)$, $\alpha_0^k \in \mathbb{R}^+$ are concentration parameters

(hyperparameters). Applying Bayes' rule, the posterior is given by $Dir(K, \alpha_j)$, $\alpha_j = (\alpha_j^1, ..., \alpha_j^K)$, where α_j^k is

$$\alpha_j^k := \alpha_0^k + \sum_{i,\ x_i \text{ in cell } j} y_i^k. \tag{2}$$

Because α_j^k counts the number of measurements which falls into the jth cell and indicate the kth category, we call this model the Semantic Counting Sensor Model (S-CSM). Given concentration parameters α_j, the mode of θ_j has the following closed form, which is also the maximum-a-posteriori estimate of θ_j:

$$\hat{\theta}_j^k = \frac{\alpha_j^k - 1}{\sum_{k=1}^K \alpha_j^k - K} \text{ and } \alpha_j^k > 1. \tag{3}$$

We also have the closed-form expected value and variance of θ_j as follows:

$$\mathbb{E}[\theta_j^k] = \frac{\alpha_j^k}{\sum_{k=1}^K \alpha_j^k} \text{ and } \mathbb{V}[\theta_j^k] = \frac{\frac{\alpha_j^k}{\sum_{k=1}^K \alpha_j^k}(1 - \frac{\alpha_j^k}{\sum_{k=1}^K \alpha_j^k})}{\sum_{k=1}^K \alpha_j^k + 1}. \tag{4}$$

We use (2) to calculate the parameters of the posterior Dirichlet distribution for cell j and given the posterior parameter α_j, the statistics of cell j can be computed by (3) and (4).

For free-class measurements, we use free-space points linearly interpolated along each sensor beam. We note that in the particular case when $K = 1$ represents the free-space class and $K = 2$ represents the occupied class, the semantic counting sensor model nicely reverts to the original counting sensor model.

However, the semantic counting sensor model inherits the traditional occupancy grid mapping limitations because the posterior parameters for each cell are only correlated with measurements that directly fall into or pass through that cell. To mitigate this shortcoming, we use BKI to convert the discrete semantic counting sensor model to a continuous model by taking into account local correlations in the map.

IV. Continuous Semantic Mapping via Bayesian Kernel Inference

Bayesian kernel inference, as introduced by Vega-Brown et al. [31], relates the extended likelihood $p(y_i|\theta_*, x_i, x_*)$ and the likelihood $p(y_i|\theta_i)$ by a smoothness constraint, where θ_* is the value of the latent variable for the query point x_*. In this framework, the maximum entropy distribution g, satisfying $D_{\text{KL}}(g\|f)$, has the form $g(y) \propto f(y)^{k(x_*,x)}$, where $D_{\text{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler Divergence (KLD), and $k(\cdot,\cdot)$ is a kernel function. Let g be the extended likelihood and f the likelihood, we define a smooth distribution over semantics as having bounded KLD between the two distributions. Given a kernel function operating on 3D spatial inputs $k : \mathcal{X} \times \mathcal{X} \to [0, 1]$, we have

$$\prod_{i=1}^N p(y_i|\theta_*, x_i, x_*) \propto \prod_{i=1}^N p(y_i|\theta_*)^{k(x_*,x_i)}. \tag{5}$$

Using Bayes' rule, we can write

$$p(\theta_*|x_*, \mathcal{D}) \propto p(\mathcal{D}|\theta_*, x_*)p(\theta_*|x_*), \tag{6}$$

and by substituting (5) into (6), we have:

$$p(\theta_*|x_*, \mathcal{D}) \propto \left[\prod_{i=1}^N p(y_i|\theta_*)^{k(x_*,x_i)}\right] p(\theta_*|x_*). \tag{7}$$

We adopt the Categorical likelihood and place a prior distribution $Dir(K, \alpha_0)$ over θ_*. Subsequently, (6) becomes:

$$\begin{aligned} p(\theta_*|x_*, \mathcal{D}) &\propto \left[\prod_{i=1}^N \left[\prod_{k=1}^K \left(\theta_*^k\right)^{y_i^k}\right]^{k(x_*,x_i)}\right] \prod_{k=1}^K \left(\theta_*^k\right)^{\alpha_0^k - 1} \\ &= \prod_{k=1}^K \left(\theta_*^k\right)^{\alpha_0^k + \sum_{i=1}^N y_i^k k(x_*,x_i) - 1}, \end{aligned} \tag{8}$$

which is proportional to the posterior $Dir(K, \alpha_*)$ where $\alpha_* = (\alpha_*^1, ..., \alpha_*^K)$ is defined as

$$\alpha_*^k := \alpha_0^k + \sum_{i=1}^N k(x_*, x_i) y_i^k. \tag{9}$$

The mode, mean, and variance for the continuous model can be computed exactly as given in (3) and (4).

Compared with (2), (9) not only considers measurements which fall into a cell but also adjacent measurements with a weighting coefficient defined by the kernel function, i.e., the distance to the query point. We note that the kernel neither needs to be positive-definite nor symmetric. To reduce the computational complexity, we choose the sparse kernel [36] as

$$k(x, x') = \begin{cases} \sigma_0 \left[\frac{1}{3}\left(2 + \cos\left(2\pi\frac{d}{l}\right)(1 - \frac{d}{l}) + \frac{1}{2\pi}\sin\left(2\pi\frac{d}{l}\right)\right)\right] & \text{if } d < l \\ 0 & \text{if } d \geq l \end{cases} \tag{10}$$

where $d = \|x - x'\|$, $l > 0$ is the length-scale, and σ_0 is kernel scale parameter (signal variance).

The derived continuous semantic model can deal with sparse and noisy sensor measurements better and allows for queries at an arbitrary resolution. In the context of semantic occupancy mapping, the query points are chosen to be the grid centroids. Thus, (9) can be used to recursively update the posterior parameters for each grid. We use a block to contain a number of grids according to the block depth, where each block is an octree of grids. For every block of test data, the corresponding training data is comprised of all portions of the new measurements that pass through the block's extended block [17], which is defined as the set of neighboring blocks with faces adjacent to the block containing the test data of interest.

Example 1 (Three-dimensional Toy Example)**.** *Figure 2 illustrates a three-dimensional toy example of the continuous semantic mapping via Bayesian kernel inference using a simulated dataset made in Gazebo, with annotated semantic labels. The simulated dataset has dimensions* $10.0 \times 7.0 \times 2.0m$*. We manually annotate the raw data into three semantic classes: ground, wall, and cylindrical obstacles. Semantic occupancy maps with resolution 0.05* m *for both S-CSM and Semantic Bayesian Kernel Inference (S-BKI) models are built using the*

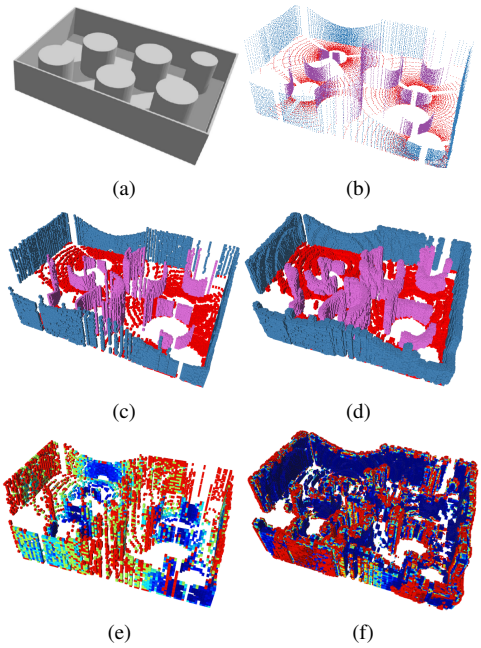

Fig. 2: 3D toy example on a simulated dataset. (a) Environment model in Gazebo. (b) Annotated point cloud raw data. (c) Semantic map of S-CSM. (d) Semantic map of S-BKI. (e) Variance map of S-CSM. (f) Variance map of S-BKI. Variance maps of two models (shown using the jet colormap) provide useful information for robotic navigation and exploration [37]. We found that Bayesian kernel inference decreases the variance of the wall by considering neighboring measurements. There are some artifacts, however, on the periphery of the wall where the variance is relatively high.

annotated point clouds as sensor measurements. The figure shows that S-CSM can reconstruct the 3D environment with correct semantic information but has a limited predictive capability where sensor coverage is sparse. The S-BKI map can interpolate the gaps in the walls due to the continuity and smoothness of Bayesian kernel inference.

V. Experimental Results

We now present experiments for evaluating semantic segmentation accuracy, occupancy prediction accuracy, and the impact of parameters using multiple real datasets. We also compare the proposed methods with state-of-the-art systems and present a qualitative evaluation using data collected with a bipedal robot. C++ implementations of the proposed methods are available open source [1], and make use of the Learning-Aided 3D Mapping Library [21], the Robot Operating System (ROS) [38], and Point Cloud Library (PCL) [39]. The parameters in Table I were manually tuned but remained fixed throughout all experiments. For baselines, we used the available open-source implementations without any modification. All experiments are conducted on an Intel i7 processor with 8 cores and 32 GB RAM.

A. KITTI Dataset

KITTI dataset with semantically labeled images contains 40 test images from sequence 05 [13], and 25 test images from sequence 15 [26] in KITTI odometry dataset. We qualitatively and quantitatively compare the mapping performance of our methods with the state-of-the-art CRF-based semantic mapping system proposed by Yang et al. [9]. However, Yang's method only predicts semantic labels on occupied voxels using a discrete occupancy grid mapping algorithm. For a fair comparison with respect to the occupancy model, we implement another baseline, BGKOctoMap-CRF [2], by replacing Yang's discrete occupancy grid map with the continuous BGKOctoMap, and then applying the same hierarchical CRF model to refine the voxel labels.

TABLE I: Kernel and Dirichlet prior hyperparameters for all experiments.

Hyperparameter	Description	Value
l	Kernel length-scale	0.3 m
σ_0	Kernel scale	0.1
α_0^k	Dirichlet prior	0.001

We adopt the same data pre-processing methods as used by Yang et al. [9]. We use ELAS [40] to generate depth maps from stereo image pairs, ORB-SLAM [41] to estimate 6DoF camera poses, and the deep network dilated CNN [42] for prior semantic label predictions. The superpixels used in Yang's CRF module and BGKOctoMap-CRF are generated by the SLIC algorithm [43]. The common parameters for occupancy mapping in all methods are set according to Yang's work: the resolution of 0.1 m, free and occupied thresholds as 0.47 and 0.6, respectively.

1) Qualitative Results: The 3D view of the semantic map built by S-BKI model is given in Fig. 1. Our approach is able to recognize and reconstruct general objects such as road, sidewalk, building, fence and vegetation. We also show the same view of the corresponding variance map of S-BKI in Fig. 1. Most of the grids on the surface have relatively low variance (cyan); the middle grids have the lowest variance (blue) where the sensor measurements are dense, while the grids on the margins of sensor scans show relatively high variance (red) where the sensor measurements are sparse. It can also be noticed that the uneven parts of the road in the semantic map have high variance, which might be caused by the discontinuity of the estimated camera poses.

We also found that a small portion of grids of the fence on the left side are misclassified as vegetation, where the corresponding variance is high. This nice property enables us to reject misclassified grids by setting a variance threshold. If the variance is too high, we can regard the state of the grid as unknown and thus build safer semantic maps for robot navigation. To compare the mapping performance, we project semantic maps onto 2D left camera views and compare with 2D ground truth images as shown in Fig. 3.

2) Quantitative Results: We follow the evaluation method given in [9] by projecting 3D semantic map onto the 2D left image plane, ignoring voxels that are too far from the camera (40 meters for all the methods), and calculating the standard metric of Intersection over Union (IoU) based on labeled

[1] https://github.com/ganlumomo/BKISemanticMapping

[2] https://github.com/zeroAska/BGKOctoMap-CRF

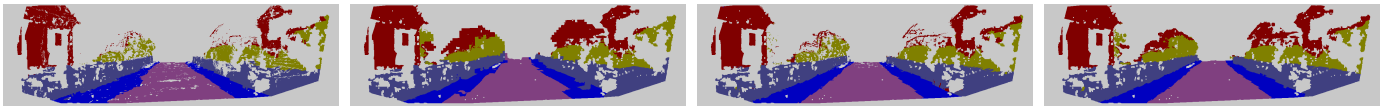

Fig. 3: Qualitative results on KITTI odometry sequence 05 dataset [13]. From left to right the figures show 2D projected images from Yang et al. [9], BGKOctoMap-CRF, S-CSM and S-BKI, respectively. The projected image from Yang's semantic map contains more gaps than other maps, compared with the ground truth image where the road, buildings, and vegetation are continuous and dense, while the projected image of S-BKI has the least holes in those regions, which resembles the ground truth better. BGKOctoMap-CRF outperforms Yang's method, in spite of the misclassification of the sidewalk to road.

TABLE II: Quantitative results on KITTI odometry sequence 05 test set [13] for 8 common semantic classes, containing 40 images.

Metric	Method	Building	Road	Vege.	Sidewalk	Car	Signate	Fence	Pole	Average
IoU Exclusive	Yang et al. [9]	86.2	91.5	**85.3**	74.1	**77.1**	16.8	78.5	28.0	67.2
	BGKOctoMap-CRF	86.1	88.0	82.3	73.6	71.9	15.5	73.8	27.7	64.9
	S-CSM	86.3	93.2	84.3	**80.0**	76.8	**25.5**	77.5	**30.1**	**69.2**
	S-BKI	**87.4**	**93.3**	84.7	79.9	76.9	18.6	**78.7**	29.2	68.6
IoU	Yang et al. [9]	32.5	70.1	45.2	55.7	39.5	13.0	46.6	18.9	40.2
	BGKOctoMap-CRF	43.5	70.9	49.4	55.5	40.2	12.7	46.4	13.9	41.6
	S-CSM	40.2	74.1	49.5	62.1	42.1	**20.3**	47.7	22.8	44.9
	S-BKI	**45.6**	**75.5**	**52.8**	**62.9**	**43.3**	14.9	**49.3**	**22.9**	**46.0**

TABLE III: Quantitative results on KITTI odometry sequence 15 test set [26] for 8 common semantic classes, containing 25 images.

Metric	Method	Building	Road	Vege.	Sidewalk	Car	Signate	Fence	Pole	Average
IoU Exclusive	Yang et al. [9]	**95.6**	90.4	**92.8**	70.0	94.4	0.1	**84.5**	49.5	72.2
	BGKOctoMap-CRF	94.7	93.8	90.2	81.1	92.9	0.0	78.0	49.7	72.5
	S-CSM	94.4	**95.4**	90.7	**84.5**	95.0	22.2	79.3	**51.6**	76.6
	S-BKI	94.6	**95.4**	90.4	84.2	**95.1**	**27.1**	79.3	51.3	**77.2**
IoU	Yang et al. [9]	32.9	85.8	59.0	**79.3**	61.0	0.9	46.8	33.9	50.0
	BGKOctoMap-CRF	**50.0**	86.6	64.1	74.9	61.0	0.0	47.5	**36.7**	52.6
	S-CSM	42.6	87.3	62.9	77.9	62.6	17.1	47.7	34.8	54.1
	S-BKI	49.3	**88.8**	**69.1**	78.2	**63.6**	**22.0**	**49.3**	**36.7**	**57.1**

ground truth left images. IoU is defined as TP/(TP+FN+FP), where T/F P/N stands for true/false positive/negative.

Yang et al. [9] *exclude* the data that has not been projected onto images (gray color in the projected images), *even when there exists corresponding ground truth data of it* (as shown in the ground truth images in Fig. 1). For a fair comparison, we follow this approach for all methods and call it *IoU Exclusive*. However, this evaluation ignores the classification error of gaps in the map, and cannot show the advantage of continuous mapping. Therefore, we compute a more rigorous *IoU* by taking all projected data except the sky class into account.

The quantitative results are given in Table II and III, where the two metrics are computed. For this experiment, the average runtime of Yang et al. is 4.41 sec/scan, BGKOctoMap-CRF is 1.10 sec/scan, S-CSM is 0.75 sec/scan, and S-BKI is 0.36 sec/scan. S-BKI has the highest IoU among almost all semantic classes compared with other maps, and S-CSM is the second-best method. We reiterate that the IoU Exclusive is not a reasonable metric for mapping performance evaluations; nevertheless, S-CSM and S-BKI still outperform the compared baselines using this metric. In the latter case, as expected, S-CSM and S-BKI perform similarly.

BGKOctoMap-CRF has a higher IoU than Yang's method because of the continuous occupancy model of BGKOctoMap. The gaps in the measurements are interpolated and CRF fills the labels from adjacent voxels. S-CSM outperforms both CRF-based methods, because even if the 3D CRF model further optimizes the grid labels, it is only post-processing pre-calculated occupied grids and, therefore, it cannot recover the correct semantic labels for misclassified occupancy or unknown grids. Specifically, even if BGKOctoMap-CRF is a continuous model for occupancy, it is not continuous for semantics and color. Thus, the predicted occupied voxels might not contain observation of semantics and color, leading to improper initialization of them for CRF potentials. In contrast, the counting sensor model uses a statistical model to infer the grid statistics. By adding the Bayesian kernel inference, S-BKI outperforms S-CSM as it can fill the gaps in the map using nearby measurements. Even for fully observed regions, by considering local correlations the map becomes more robust to noisy measurement.

B. *SemanticKITTI Dataset*

SemanticKITTI [30] is a large-scale dataset based on the KITTI odometry dataset. It provides dense annotations for each LiDAR scan of 22 sequences including camera poses estimated from a surfel-based SLAM approach (SuMa) [44]. The input data of this dataset is collected by a Velodyne HDL-64E laser scanner. The semantic measurements are generated by RangeNet++ [45], which is a state-of-the-art LiDAR-only semantic segmentation deep neural network. To investigate mapping performance on noisy data, we choose two backends provided in RangeNet++: the best-performing one, Darknet53-kNN, and SqueezeSegV2-kNN which has lower performance. All maps are built at a resolution of 0.1 m and without any pre-processing of the input data.

For evaluation, we use all sequences in SemanticKITTI. For training (00-07, 09-10) and validation (08) sequences, we compare the 3D predictions with ground truth labels. To obtain the IoU metrics for test (11-21) sequences, we submitted our results to the official evaluation server which are shown on the multi-scan leaderboard[3]. As our method is for static environments, we cannot differentiate between static and dynamic objects. For static semantic classes, we outperforms Darknet53-kNN for 18 out of 19 classes on test sequences.

Quantitative results on all sequences are given in Table IV. For this experiment, the average runtime of S-CSM is 9.48 sec/scan, S-BKI is 1.67 sec/scan. For all sequences, our semantic mapping methods can improve the prior segmentation IoU by fusing multiple scans. We note that S-BKI consistently outperforms S-CSM in almost all semantic classes, which shows the advantage of Bayesian kernel inference and continuous semantic maps. When S-CSM does outperform S-BKI, the IoUs are close to each other. Moreover, the mapping improvement over SqueezeSegV2-kNN is much higher than Darknet53-kNN, which shows our methods can deal with noisy input data.

[3] https://competitions.codalab.org/competitions/20331#results (ganlumm)

TABLE IV: Mean IoU on SemanticKITTI dataset sequence 00-21 [30] for 19 semantic classes. SqueezeSegV2-kNN (Sq.-kNN). Darknet53-kNN (Da.-kNN). Training (00-07, 09-10), Validation (8), Test (11-21).

Seq.	Method	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign	Average
Training	Sq.-kNN	88.2	14.4	45.7	67.3	60.9	33.3	58.7	63.1	92.6	62.0	81.3	49.2	77.3	63.6	76.7	34.5	71.5	32.8	49.5	59.1
	S-CSM (w/ Sq.-kNN)	92.6	21.6	62.2	73.1	70.6	44.1	80.3	67.4	94.3	70.9	85.1	52.3	82.0	69.1	81.4	47.8	75.4	50.8	65.0	67.7
	S-BKI (w/ Sq.-kNN)	93.5	29.1	73.9	82.0	77.0	54.6	87.2	73.7	93.8	73.6	84.2	55.7	83.8	70.1	82.8	53.9	75.9	54.6	70.4	72.1
	Da.-kNN	94.7	42.3	81.8	83.4	69.4	69.4	72.5	53.7	96.7	88.6	92.6	82.1	95.4	85.0	92.0	71.0	88.3	70.6	82.4	79.6
	S-CSM (w/ Da.-kNN)	96.0	48.6	88.3	84.5	71.4	77.3	83.6	54.3	**96.8**	89.7	**93.3**	84.2	96.6	86.7	93.5	79.2	90.0	80.0	88.9	83.3
	S-BKI (w/ Da.-kNN)	**96.9**	**53.2**	**90.9**	**85.9**	**73.3**	**83.5**	**88.4**	**59.8**	**96.8**	**89.9**	93.1	**85.4**	**97.3**	**87.4**	**94.2**	**81.3**	**90.9**	**82.0**	**90.6**	**85.3**
Validation	Sq.-kNN	86.7	14.4	24.6	21.0	23.3	23.5	40.9	0.0	90.1	32.4	74.8	**1.2**	79.6	42.7	79.2	36.5	71.1	28.3	24.8	41.8
	S-CSM (w/ Sq.-kNN)	90.5	23.0	34.9	26.8	29.1	32.4	49.4	0.0	92.6	38.7	79.0	1.1	84.6	51.6	83.3	48.3	72.9	44.1	31.6	48.1
	S-BKI (w/ Sq.-kNN)	92.3	30.0	39.7	29.3	**32.1**	38.8	54.7	0.0	92.9	40.9	79.9	1.1	86.6	54.6	84.9	52.3	74.2	47.9	34.7	50.9
	Da.-kNN	91.0	25.0	47.1	40.7	25.5	45.2	62.9	0.0	93.8	46.5	81.9	0.2	85.8	54.2	84.2	52.9	72.7	53.2	40.0	52.8
	S-CSM (w/ Da.-kNN)	92.6	32.5	54.9	43.4	26.2	51.3	69.2	0.0	**94.6**	49.2	**84.0**	0.1	87.9	58.4	85.8	59.9	73.3	61.7	43.0	56.2
	S-BKI (w/ Da.-kNN)	**93.5**	**33.5**	**57.3**	**44.5**	27.2	**52.9**	**72.1**	0.0	94.4	**49.6**	**84.0**	0.0	**88.7**	**59.6**	**86.9**	**62.5**	**75.3**	**63.6**	**45.1**	**57.4**
Test	Da.-kNN	82.4	26.0	34.6	21.6	18.3	6.7	2.7	**0.5**	91.8	65.0	75.1	27.7	87.4	58.6	80.5	55.1	64.8	47.9	55.9	47.5
	S-BKI (w/ Da.-kNN)	**83.8**	**30.6**	**43.0**	**26.0**	**19.6**	**8.5**	**3.4**	0.0	**92.6**	**65.3**	**77.4**	**30.1**	**89.7**	**63.7**	**83.4**	**64.3**	**67.4**	**58.6**	**67.1**	**51.3**

TABLE V: Comparison of map quality using the Area Under ROC Curve (AUC) and runtime of the four methods on the example shown in Fig. 4.

Method	OctoMap	BKIOctoMap	S-CSM	S-BKI
AUC	0.7226	**0.7801**	0.7274	**0.7801**
Runtime (s)	252.32	73.30	480.44	**68.17**

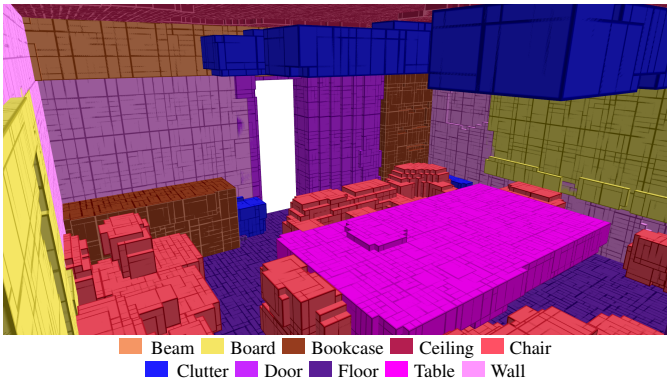

Fig. 4: S-BKI map of a conference room in Area 3 of Stanford 2D-3D-Semantics Dataset [46].

C. Occupancy Evaluation

To support the claim that S-BKI is a semantic occupancy mapping method, we evaluate the accuracy of occupancy prediction of S-CSM, S-BKI, OctoMap and BGKOctoMap. The experiment is performed using a conference room in Area 3 of Stanford 2D-3D-Semantics Dataset [46], as ground-truth occupancy values are provided. For S-CSM and S-BKI, we use the annotated point clouds to build the semantic occupancy maps, and the same point clouds without semantics for OctoMap and BGKOctoMap. The semantic map built by S-BKI is shown in Fig. 4. Comparisons of map quality and runtime of the four methods are given in Table V. For S-CSM and S-BKI, the probability of occupancy is computed as the sum of all probabilities of valid semantic classes. Among all methods, S-BKI and BGKOctoMap have the highest (identical) performance, which shows that S-BKI reduces to BGKOctoMap when only occupancy is of interest, not only theoretically, but also experimentally. S-CSM is slower than S-BKI because the block depth is set to one, thus S-CSM has more blocks to be computed.

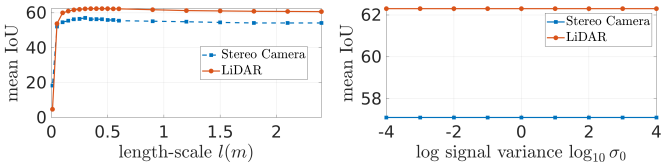

Fig. 5: Impact of parameters on mapping performance for KITTI dataset sequence 15 [26] (stereo camera) and SemanticKITTI dataset sequence 04 [30] (LiDAR). Only one parameter at a time is varied while the others are kept at the values in Table I. Both figures show reasonable robustness to the parameter variations.

D. Impact of Parameters

We empirically study the sensitivity of S-BKI mapping to the kernel length-scale and signal variance. The experiments are conducted using KITTI dataset sequence 15 [26] for stereo camera and SemanticKITTI datase sequence 04 [30] for LiDAR data. All other parameters are fixed to the values indicated in Table I. In Fig. 5, we plot the kernel length-scale l and signal variance σ_0 against the mean IoU metrics. The influence of the kernel length-scale on mapping performance for both stereo camera and LiDAR data is similar: the mean IoU increases rapidly as the length-scale varies from 0.01 to 0.1, gradually increases to a peak value, and then drops gradually as the length-scale increases. S-BKI achieves the best performance when $l = 0.3$ for stereo camera data and $l = 0.4$ for LiDAR data. This is reasonable because LiDAR data is sparser than stereo camera data and longer distance should be considered. The mapping performance is insensitive to signal variance over a large scale; this is because we use the same signal variance for all semantic classes. To see an effect, one would need to allow signal variance to vary from class-to-class.

E. Experimental Results on a Cassie Bipedal Robot

We test our mapping methods on data collected using the bipedal robot Cassie Blue shown in Fig 6. To obtain semantic measurements, we manually annotated 1194 training images and 457 validation images from the NCLT dataset [47].

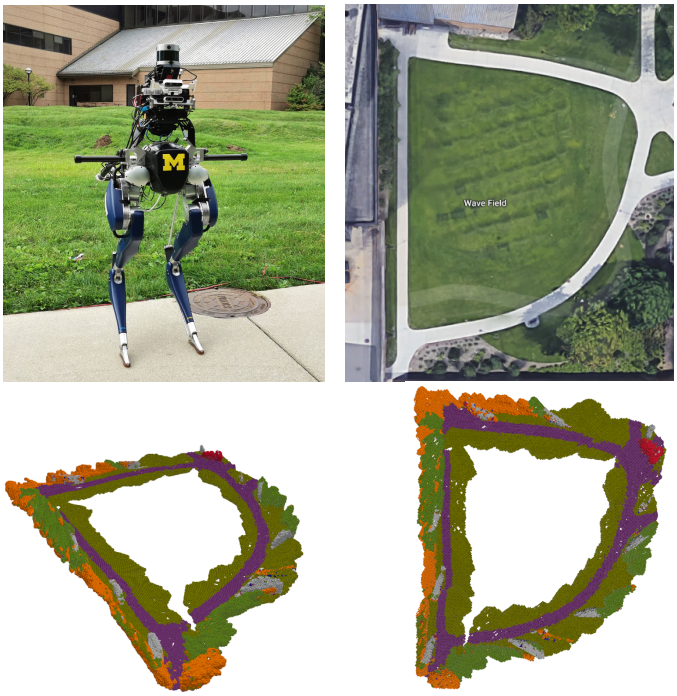

Fig. 6: Top Left: Cassie Blue has a custom designed torso on which is mounted an Intel RealSense depth camera capable of providing both RGB images and corresponding organized point clouds in outdoor environments. Top right: Google satellite map of the Wave Field of the University of Michigan - North Campus. Bottom: From left to right are the 3D and 2D views of S-BKI map. While the robot is navigating along the sidewalk, S-CSM produces discontinuous semantic maps from sparse sensor measurements, which may cause the robot's planner to regard the gaps in the map as unwalkable areas, a practical problem when we conduct autonomous walking experiments with Cassie Blue (a video of the experiment is available at https://www.youtube.com/watch?v=uFyT8zCg1Kk&t=3s). S-BKI model produces a continuous and smooth map, where gaps are assigned with labels inferred from local correlations in the map.

The NCLT dataset was selected because it shares a similar environmental domain as the Wave Field data, which includes *background, water, road, sidewalk, terrain, building, vegetation, car, person, bike, pole, stair, traffic sign and sky* for a total of 14 classes. We used these images to fine-tune a modified 2D segmentation network MobileNet [48] with a pre-trained model on the ImageNet dataset [49] for efficiency. The fine-tuned network segments the RGB images, and then we can directly label the organized point clouds.

The qualitative results are given in Fig. 6. To show the mapping performance of our methods on sparse data, we downsample the point clouds per scan to a resolution of 0.2 m, and build a semantic occupancy map with a resolution of 0.1 m. S-BKI runs at about 2 Hz. The mapping drift after one full round of the Wave Field is because of the odometry system [50] instead of SLAM used in the experiment.

VI. Discussions and Limitations

In practice, semantic measurements do not necessarily come in the form of a one-hot vector, but rather a pseudo-probability vector obtained from the softmax output of a classifier. Taking the max rather than the softmax results in the current formulation. Taking the softmax, on the other hand, results in other models corresponding to a set of model-averaging techniques (i.e., the linear opinion pooling and Nadaraya-Watson kernel-regression) that are similar, but not identical, to the Bayesian model presented in Sec. III.

There are still several limitations to this work. First, the length-scale of the kernel function trades off predictive ability and classification accuracy. When the length-scale is large, the model can extrapolate large-scale trends in data, and thus be more predictive; however, the classification accuracy may drop for small objects in the environment. In the current approach, we manually tune the length-scale and use the same scale everywhere, independent of the class. Optimizing the hyperparameters in a Bayesian framework can be helpful. In addition, varying the length-scale and signal variance based on geometric features and semantic properties may further improve semantic mapping performance. Secondly, the memory and space storage for large-scale mapping is another limitation. We currently store the entire semantic map in computer memory without any pruning. However, with the current test-data octrees data structure, even when storing the map after pruning, the save in memory consumption is not substantial. How to compress the continuous semantic maps is an interesting future research direction. Finally, the current semantic map is for static environments, differentiating between static and dynamic semantic labels is also an interesting future work.

VII. Conclusion

In this paper, we extended the counting sensor model for occupancy grid mapping to a semantic counting sensor model for semantic occupancy mapping. To relax the independent-grid assumption in occupancy grid mapping, we used a Bayesian spatial kernel inference to generalize the semantic counting sensor model to continuous semantic mapping. Extensive experimental results show the proposed methods work with both dense stereo camera and LiDAR data. We improved the mapping performance over the state-of-the-art semantic mapping system using the KITTI dataset, and increased the segmentation accuracy over a 3D deep neural network with kNN processing using the SemanticKITTI dataset. We labeled the NCLT dataset and collected data using Cassie Blue biped robot to further evaluate the mapping performance in real world experiments. The S-BKI model consistently outperforms S-CSM, which shows the advantage of using Bayesian kernel inference in continuous mapping.

Acknowledgment

This article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. The authors would like to thank Yukai Gong for the development of the feedback controller utilized in the Cassie experiments as well as Bruce Huang, Zhenyu Gan, Omar Harib, Eva Mungai, and Grant Gibson for their help in collecting experimental data.

References

[1] H. P. Moravec and A. Elfes, "High resolution maps from wide angle sonar," in *Proc. IEEE Int. Conf. Robot. and Automation*, vol. 2. IEEE, 1985, pp. 116–121.

[2] A. Elfes, "Sonar-based real-world mapping and navigation," *IEEE J. Robot. Autom.*, vol. 3, no. 3, pp. 249–265, 1987.

[3] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: an efficient probabilistic 3D mapping framework based on octrees," *Auton. Robot.*, vol. 34, no. 3, pp. 189–206, 2013.

[4] A. Nüchter and J. Hertzberg, "Towards semantic maps for mobile robots," *Robot. and Auton. Syst.*, vol. 56, no. 11, pp. 915–926, 2008.

[5] D. F. Wolf and G. S. Sukhatme, "Semantic mapping using mobile robots," *IEEE Trans. Robot.*, vol. 24, no. 2, pp. 245–258, 2008.

[6] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robot. and Auton. Syst.*, vol. 66, pp. 86–103, 2015.

[7] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2015, pp. 75–82.

[8] S. Sengupta and P. Sturgess, "Semantic octree: Unifying recognition, reconstruction and representation via an octree constrained higher order MRF," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2015, pp. 1874–1879.

[9] S. Yang, Y. Huang, and S. Scherer, "Semantic 3D occupancy mapping through efficient high order CRFs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.* IEEE, 2017, pp. 590–597.

[10] J. Stückler, N. Biresev, and S. Behnke, "Semantic mapping using object-class segmentation of RGB-D images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.* IEEE, 2012, pp. 3005–3010.

[11] B.-S. Kim, P. Kohli, and S. Savarese, "3D scene understanding by voxel-CRF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1425–1432.

[12] Z. Zhao and X. Chen, "Building 3D semantic maps for mobile robots using RGB-D camera," *Intell. Service Robot.*, vol. 9, no. 4, pp. 297–309, 2016.

[13] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg, "Joint semantic segmentation and 3D reconstruction from monocular video," in *Proc. European Conf. Comput. Vis.* Springer, 2014, pp. 703–718.

[14] S. T. O'Callaghan and F. T. Ramos, "Gaussian process occupancy maps," *Int. J. Robot. Res.*, vol. 31, no. 1, pp. 42–62, 2012.

[15] S. Kim and J. Kim, "GPmap: A unified framework for robotic mapping based on sparse Gaussian processes," in *Field Service Robot.* Springer, 2015, pp. 319–332.

[16] M. Ghaffari Jadidi, J. Valls Miró, R. Valencia, and J. Andrade-Cetto, "Exploration on continuous Gaussian process frontier maps," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2014, pp. 6077–6082.

[17] J. Wang and B. Englot, "Fast, accurate Gaussian process occupancy maps via test-data octrees and nested Bayesian fusion," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2016, pp. 1003–1010.

[18] F. Ramos and L. Ott, "Hilbert maps: Scalable continuous occupancy mapping with stochastic gradient descent," *Int. J. Robot. Res.*, vol. 35, no. 14, pp. 1717–1730, 2016.

[19] M. Ghaffari Jadidi, J. Valls Miro, and G. Dissanayake, "Gaussian processes autonomous mapping and exploration for range-sensing mobile robots," *Auton. Robot.*, vol. 42, no. 2, pp. 273–290, 2018.

[20] K. Doherty, J. Wang, and B. Englot, "Bayesian generalized kernel inference for occupancy map prediction," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2017, pp. 3118–3124.

[21] K. Doherty, T. Shan, J. Wang, and B. Englot, "Learning-aided 3-D occupancy mapping with Bayesian generalized kernel inference," *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 953–966, 2019.

[22] M. Ghaffari Jadidi, L. Gan, S. A. Parkison, J. Li, and R. M. Eustice, "Gaussian processes semantic map representation," *arXiv preprint arXiv:1707.01532*, 2017.

[23] L. Gan, M. Ghaffari Jadidi, S. A. Parkison, and R. M. Eustice, "Sparse Bayesian inference for dense semantic mapping," *arXiv preprint arXiv:1709.07973*, 2017.

[24] D. Hahnel, R. Triebel, W. Burgard, and S. Thrun, "Map building with mobile robots in dynamic environments," in *Proc. IEEE Int. Conf. Robot. and Automation*, vol. 2. IEEE, 2003, pp. 1557–1563.

[25] H. He and B. Upcroft, "Nonparametric semantic segmentation for 3D street scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.* IEEE, 2013, pp. 3697–3703.

[26] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr, "Urban 3D semantic modelling using stereo vision," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2013, pp. 580–585.

[27] J. P. C. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. S. Torr, "Mesh based semantic modelling for indoor and outdoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2067–2074.

[28] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2017, pp. 4628–4635.

[29] X. Chen, A. Milioto, E. Palazzolo, P. Giguère, J. Behley, and C. Stachniss, "SuMa++: Efficient LiDAR-based semantic SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.* IEEE, 2019.

[30] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9297–9307.

[31] W. Vega-Brown, M. Doniec, and N. Roy, "Nonparametric Bayesian inference on multivariate exponential families," in *Proc. Advances Neural Inform. Process. Syst. Conf.*, 2014, pp. 2546–2554.

[32] V. Peretroukhin, W. Vega-Brown, N. Roy, and J. Kelly, "PROBE-GK: Predictive robust estimation using generalized kernels," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2016, pp. 817–824.

[33] C. Richter, W. Vega-Brown, and N. Roy, "Bayesian learning for safe high-speed navigation in unknown environments," in *Robot. Res.* Springer, 2018, pp. 325–341.

[34] T. Shan, J. Wang, B. Englot, and K. Doherty, "Bayesian generalized kernel inference for terrain traversability mapping," in *Proc. Conf. Robot Learning*, 2018, pp. 829–838.

[35] D. Hähnel, "Mapping with mobile robots," Ph.D. dissertation, University of Freiburg, Freiburg im Breisgau, Germany, 2005.

[36] A. Melkumyan and F. Ramos, "A sparse covariance function for exact Gaussian process inference in large datasets," in *Proc. Int. Joint Conf. Artif. Intell.*, 2009, pp. 1936–1942.

[37] M. Ghaffari Jadidi, J. Valls Miro, and G. Dissanayake, "Sampling-based incremental information gathering with applications to robotic exploration and environmental monitoring," *Int. J. Robot. Res.*, vol. 38, no. 6, pp. 658–685, 2019.

[38] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "ROS: an open-source Robot Operating System," in *ICRA workshop on open source software*, vol. 3, no. 3.2, 2009.

[39] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *Proc. IEEE Int. Conf. Robot. and Automation*. IEEE, 2011.

[40] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Proc. Asian Conf. Comput. Vis.* Springer, 2010, pp. 25–38.

[41] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[42] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[43] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2281, 2012.

[44] J. Behley and C. Stachniss, "Efficient surfel-based SLAM using 3D laser range data in urban environments," in *Proc. Robot.: Sci. Syst. Conf.*, 2018.

[45] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet++: Fast and accurate LiDAR semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2019.

[46] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.

[47] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan North Campus long-term vision and lidar dataset," *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1023–1035, 2016.

[48] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, and M. Jagersand, "RtSeg: Real-time semantic segmentation comparative study," in *Proc. Int. Conf. Image Process.* IEEE, 2018, pp. 1603–1607.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2009, pp. 248–255.

[50] R. Hartley, M. G. Jadidi, J. Grizzle, and R. M. Eustice, "Contact-aided invariant extended Kalman filtering for legged robot state estimation," in *Proc. Robot.: Sci. Syst. Conf.*, June 2018.