Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-20 (2021) 820–825

3D Semantic Mapping: a Benchmark and Baseline Method

Guoxiang Zhang *,⋆ **Peng Wang** **,⋆ **Ning Chen** ***
YangQuan Chen ****

* *Mechatronics, Embedded Systems and Automation (MESA) Lab, University of California, Merced, Merced, CA 95348 USA (e-mail: gzhang8@ucmerced.edu).*
** *Department of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, 210037, P. R. China (e-mail: pwang41@ucmerced.edu)*
*** *Department of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, 210037, P. R. China (e-mail: chenning@njfu.com.cn)*
**** *Mechatronics, Embedded Systems and Automation (MESA) Lab, University of California, Merced, Merced, CA 95348 USA (e-mail: ychen53@ucmerced.edu).*

Abstract: In this paper, a simple and effective real-time 3D semantic mapping method is proposed. The proposed method take per-frame bounding box detections and sensor (camera) extrinsic transformation estimates as inputs and produces a set of static 3D bounding boxes in world coordinate system as 3D semantic mapping results. Each object has a Kalman filter as its state estimator and intersection over union of 3D bounding boxes is used for data association. To evaluate the proposed method, a new benchmark is derived from the KITTI object tracking evaluation. Ground-truth semantic maps are constructed based on oxts data and labeled 3D bounding boxes of KITTI. Three novel semantic map-centered metrics: DAOD, AAOD, and PRVO are proposed. Experiments are conducted to evaluate the proposed method. Experiments show that the metric and benchmarking dataset can serve as a new benchmark platform for easier comparison of new methods.

Keywords: semantic mapping, 3D perception, autonomous driving

1. INTRODUCTION

Semantic maps are important data abstraction from 3D mapping results. With higher-level semantic information, robots and autonomous vehicles can get better environmental awareness and do longer-term path and motion planning. It may also make the storing, transmitting, and retrieving of 3D mapping results more feasible, because semantic information usually has better long-term stability with smaller data footage.

At present, the concept of semantic mapping is not crystal clear yet Bavle et al. (2020) because it is still under active exploration. Some research works define semantic mapping as per 3D point/surfel/voxel labeling on dense mapping results McCormac et al. (2017); Nakajima et al. (2018); Grinvald et al. (2019), which, we believe, extends the idea of image segmentation to 3D space. Some others choose to build 3D point cloud maps with only traffic lines, and signs remained Qin et al. (2020). In this work, we focus on mapping 3D objects in the form of 3D bounding boxes for 3D semantic maps that are easy to store and transmit. Another motivation is that 3D object detection has shown promising performance improvements recently. Thus 3D bounding box input is increasingly accessible.

We list contributions of this paper:

(1) A simple and effective real-time 3D semantic mapping method is proposed. The proposed method takes per-frame bounding box detections and sensor (camera) extrinsic transformation estimates as inputs and produces a set of static 3D bounding boxes in a world coordinate system as 3D semantic mapping results.
(2) A new benchmark is derived from the KITTI object tracking evaluation, since KITTI has no official ground truth semantic maps. In the new benchmark, ground-truth semantic maps are constructed based on GPS-IMU data and labeled 3D bounding boxes of KITTI.
(3) Three novel semantic map-centered metrics are proposed for better evaluation of semantic mapping methods.

2. RELATED WORK

3D Object Detection. The research of 2D object detection has been very mature. Representative works can be divided into two main categories: Region Proposal

⋆ Both authors contributed equally to this work.

10.1016/j.ifacol.2021.11.273

Network (RPN) approaches, such as Faster R-CNN Pargi et al. (2019) and Mask R-CNN Zhou et al. (2019). Single Shot MultiBox Detector (SSD) approaches like YOLO v1-YOLO v5. Compared with 2D object detection, there is a new requirement for 3D object detection. The 3D bounding box has three more angles: pitch, yaw, and roll, in addition to the position and size. In Shi et al. (2019), PointRCNN for 3D object detection from raw point clouds is proposed. The whole framework is composed of two stages: stage-1 for the bottom-up 3D proposal generation and stage-2 for refining proposals in the canonical coordinates to obtain the final detection results. Instead of generating proposals from RGB images or projecting point clouds to bird's views, the stage-1 sub-network directly generates a small number of high-quality 3D proposals from point clouds in a bottom-up manner via segmenting the point cloud of the whole scene into foreground points and background. The stage-2 sub-network transforms the pooled points of each proposal to canonical coordinates to learn better local spatial features, which is combined with global semantic features of each point learned in stage-1 for accurate box refinement and confidence prediction.

3D Multi-Object Tracking. In Weng et al. (2020), a 3D MOT system is proposed. An off-the-shelf 3D object detector is used to obtain oriented 3D bounding boxes from the LiDAR point cloud. Then, a combination of the 3D Kalman filter and Hungarian algorithm is used for state estimation and data association. Although it is a straightforward combination of standard methods, good results are observed on the KITTI dataset.

3. SEMANTIC MAPPING BENCHMARK DATASET

It requires a large number of human efforts to annotate ground-truth labels for creating ground-truth semantic maps. So, instead of labeling new datasets, we turn to existing datasets and explore ways to convert feasible ones for semantic mapping evaluation.

Although there are several datasets with 3D bounding box annotations, as shown in Table 1, the purpose of these annotations is to provide ground truth for 3D object detection and 3D object tracking algorithms in autonomous driving scenes. For example, although nuScenes Caesar et al. (2020) dataset has the most labels, most of them are dynamic vehicles in highway scenes. For semantic mapping purposes, a large number of static objects are required. After comparison, we found that the multi-object tracking benchmark in the KITTI dataset meets the requirements and can be transformed for semantic mapping evaluation. This benchmark in KITTI has multiple data sequences with color stereo camera images, gray stereo camera images, LiDAR point clouds, GPS-IMU data together with ground-truth 3D bounding boxes for objects, and ground-truth data association among these objects.

Ground-truth camera trajectory. With the available GPS-IMU data, ground truth trajectory T_{wi} in the GPS-IMU coordinate frame can be derived. In the KITTI dataset, extrinsic parameters between sensors are also given. There are two transformation matrices that are useful for transforming GPS-IMU trajectory to camera trajectory. One is from the GPS-IMU coordinate to velodney coordinate T_{vi}. The another is from velodney coordinate to camera coordinate T_{cv}. Thus, the transformation from the GPU-IMU coordinate to the camera 0 frame is $T_{ci} = T_{P0} \cdot T_{cv} \cdot T_{vi}$. With this transformation, we can get the ground-truth trajectory of camera 0 with:

$$T_{wc} = T_{ci} \cdot T_{i0w} \cdot T_{wi} \cdot T_{ic}, \tag{1}$$

where T_{wc} is the transformation matrix from the camera coordinate to the world coordinate system.

Ground-truth semantic map. Considering that an object has several different 3D bounding box labels across multiple frames, and these boxes will not overlap perfectly due to a number of error factors. To address this problem, we fuse these boxes of the same object in the world coordinate system. First, the center and size of a 3D bounding box is computed by

$$C_w = \frac{\sum_{i=1}^{N_b} C_{wi}}{N_b}, \tag{2}$$

where C_w is x or y or z, which represents the center of a object 3D bounding box. N_b is the number of these boxes of the i-th object belongs to $(0, N_b)$.

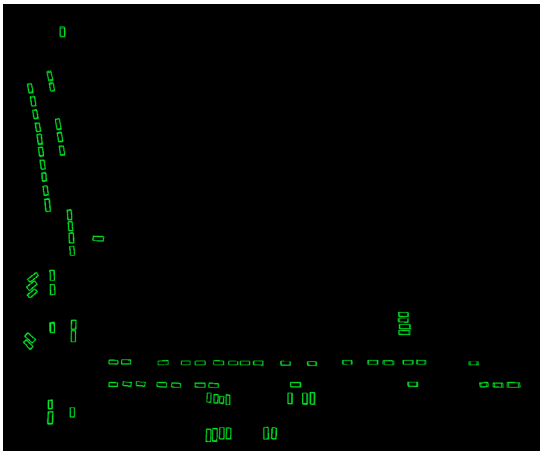
For the orientation of these bounding boxes, obviously, we can not use the same method. In the KITTI dataset, the θ is used to represent the orientation of the object. It can be converted to a rotation matrix

$$R_{cb}(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix}. \tag{3}$$

Then, orientation R_{wb} of the 3D bounding box in the world coordinate system can be computed by

$$R_{wb} = R_{wc} R_{cb}, \tag{4}$$

By converting the rotation matrix R_{wb} to the Euler angle representation, we have three Euler angles. Then, we can use the same method as in equation (2).



(a) sequence 01

(b) sequence 11

Fig. 1. Visualization of our generated ground-truth semantic maps. Each green box is for a parked car.

Table 1. Datasets with ground-truth 3D bounding boxes available.

Dataset	Year	Sequence count	RGB frame #	LiDAR frame #	3D bounding box #
KITTI	2012	22	15K	15K	200K
AS lidar	2018	-	0	20K	475K
ApolloScape	2018	-	144K	0	70K
H3D	2019	160	83K	27K	1.1M
nuScenes	2019	1K	1.4M	400K	1.4M

Table 2. Information of the generated ground-truth semantic maps. Storage is caculated when stored in 64 bit floating point numbers.

Sequence	Static object number	Odometry (m)	Storage (KB)
01	81	361	5.8
07	48	352	3.5
09	70	614	5.0
11	40	224	2.9
14	12	61	0.9

Finally, we get the 3D bounding boxes ground truth of each object in the world coordinate system. Two of these ground-truth semantic maps are visualized in Fig. 1. For semantic mapping evaluation, we find five suitable sequences: 01, 07, 09, 11, 14 from the KITTI multi-object tracking benchmark. The information of these maps is shown in Table 2.

4. SEMANTIC MAPPING METHOD

The proposed semantic mapping method takes camera tracking and 3D object detection as input to process then generate 3D bounding boxes in world frames. This process will then fuse observation of objects across different frames into a map of objects in the form of 3D bounding boxes. The core of the mapping method is Kalman filter and ROI with Huganrian method as data association.

We define states of a Kalman filter for object fusion as $x, y, z, \varphi, \theta, \psi, l, w, h, v_x, v_y, v_z$, which consists of bounding box location center x, y, z, the three angle that represents the direction of a bounding box, length l, width w, height h, and speed of the movement of the bounding box v_x, v_y, v_z. For the defined state, the state transition function is

$$\begin{cases} x^{k+1} = x^k + v_x^k + \sigma_x, \\ y^{k+1} = y^k + v_y^k + \sigma_y, \\ z^{k+1} = z^k + v_z^k + \sigma_z, \\ \varphi^{k+1} = \varphi^k + \sigma_\varphi, \\ \theta^{k+1} = \theta^k + \sigma_\theta, \\ \psi^{k+1} = \psi^k + \sigma_\psi, \\ l^{k+1} = l^k + \sigma_l, \\ w^{k+1} = w^k + \sigma_w, \\ h^{k+1} = h^k + \sigma_h, \\ v_x^{k+1} = v_x^k + \sigma_{vx} \\ v_y^{k+1} = v_y^k + \sigma_{vy} \\ v_z^{k+1} = v_z^k + \sigma_{vz} \end{cases} \tag{5}$$

where each σ represents state transition noise for a state dimension. The observation functions are more simpler since all the state except the velocities are directly observable.

$$\begin{bmatrix} x_o^k \\ y_o^k \\ z_o^k \\ \varphi_o^k \\ \theta_o^k \\ \psi_o^k \\ l_o^k \\ w_o^k \\ h_o^k \end{bmatrix} = \begin{bmatrix} 1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0&0 \end{bmatrix} \begin{bmatrix} x^k \\ y^k \\ z^k \\ \varphi^k \\ \theta^k \\ \psi^k \\ l^k \\ w^k \\ h^k \\ v_x^k \\ v_y^k \\ v_z^k \end{bmatrix} + \boldsymbol{\sigma}_{obs} \tag{6}$$

where underscript o denotes observation of the the data dimension. Note that the abstracted observation is in world frame thus the real observation need to be first transformed to the world frame. Given the detection bounding boxes set $\mathbb{B}_c$, where each bound box denoted as B_c and camera pose Twc_i. $\mathbb{B}_c$ are transformed to world frame, denoted as

$$\mathbb{B}_w = Twc_i(\mathbb{B}_c), \tag{7}$$

where $Twc_i(\cdot)$ is a function that transfer bounding boxes with the transformation Twc_i. Given the abstracted observation $\mathbb{B}_w$ and bounding box predictions $\overline{\mathbb{B}_w}$, intersection over union (IoU) of two set of bounding boxes are calculated and fed into the Hungarian algorithm to find data association between them. If a $B_w \in \mathbb{B}_w$ is matched with a predicted bounding boxes, the observation will be used to update the corresponding Kalman filter. Otherwise, it will be used to initialize a new object with new Kalman filter on the semantic map.

4.1 Object state transition

There is noise in object detection results that are fed into the pipeline. The false positive object detections would cause false objects in mapping results if not handled. To deal with issue, we propose adding object states to an object in a map. The states are 1) unstable object, which is the entry state when a new object is created for the map. An object will stay in unstable state if the object is observed less than h times. 2) dynamic object, which is an object that is moving. Any object that ever moved with velocity greater than v_t will have this state. 3) static activate, which is an object with velocity smaller than v_t all the time and observed at least in k frames. 4) static inactivate, which is the destination of static activate object when not observed once in the last k frames. 5) inactivate state, which will take unstable and dynamic objects that are not observed once in the last k frames. The transition between the states is shown in Fig. 2. With these states, we take all the static objects as the mapping results.

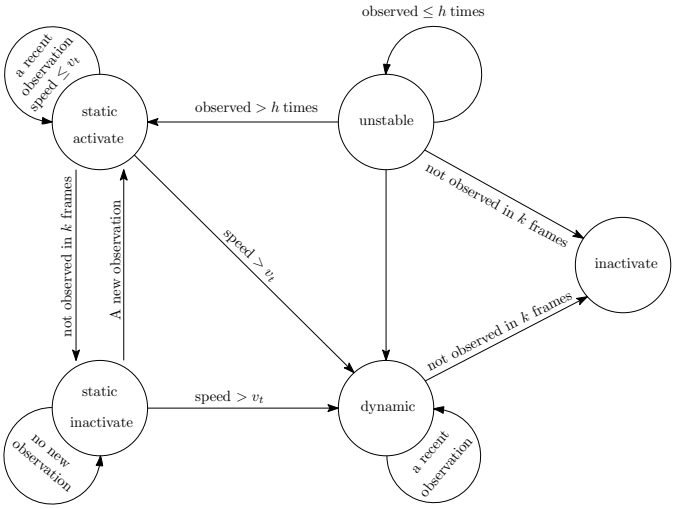


Fig. 2. State transition finite state machine

5. EVALUATION OF SEMANTIC MAPPING

To evaluate semantic mapping methods, existing metrics: intersection over union and precision-recall are adopted. In addition to that, three additional map-focused metrics are proposed to better evaluate the performance of semantic mapping methods. Two of them, as in section 5.3, focus on full-map absolute distances between matched objects, while the other one, as in section 5.5, focuses on frame-based relative error in virtual observations.

We run simple baseline experiments to evaluate the performance of the proposed method on the aforementioned KITTI sequences with the metrics.

5.1 Visual evaluation

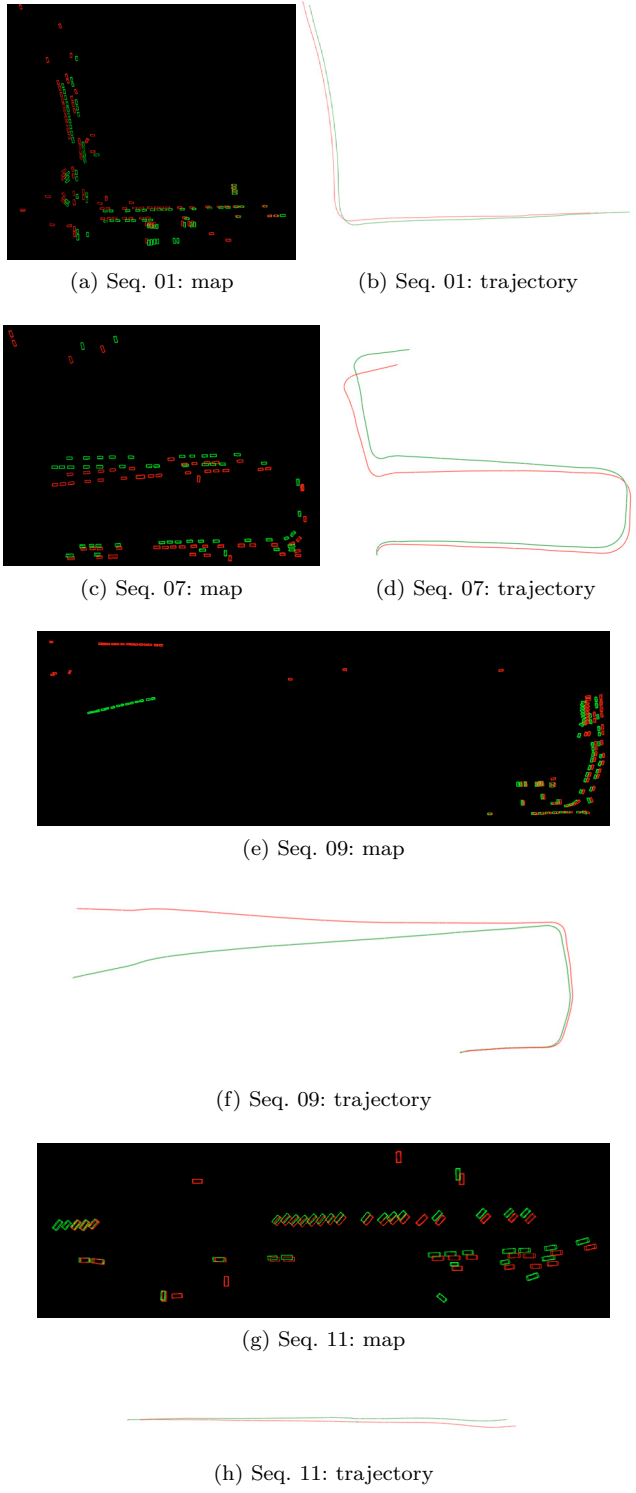
Semantic mapping results of the proposed baseline method are visualized and reported in Fig. 3. To better understand mapping performance, the input camera trajectories are also visualized comparing ground-truth camera trajectories. In the figures, we can see that the estimated 3D bounding boxes in the estimated maps are accurate relative to estimated camera trajectories. Most of the objects presented in the ground-truth maps are detected in the estimated maps. Red: estimated. Green: ground-truth.

5.2 Precision and recall

For object detection, precision and recall are usually used to analyze the effectiveness of the detection methods quantitatively. The true positive (TP) detections are the detections that agree with ground-truth labels with higher than threshold IoU values. False positive (FP) represents the bounding boxes that are detected by an algorithm but do not appear in the ground truth. False negative (FN) represents the number of ground-truth bounding boxes that are missing from detection results. Precision and recall then can be calculated as in (8) and (9), respectively.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{9}$$



(a) Seq. 01: map

(b) Seq. 01: trajectory

(c) Seq. 07: map

(d) Seq. 07: trajectory

(e) Seq. 09: map

(f) Seq. 09: trajectory

(g) Seq. 11: map

(h) Seq. 11: trajectory

Fig. 3. Visualization of estimated ground-truth semantic map and input camera trajectory for sequence 01, 07, 09, 11. Red: estimated result. Green: ground-truth data. We can see that the estimated 3D bounding boxes in the estimated map are accurate relative to estimated camera trajectory. Most of objects presented in the ground-truth map are detected in the estimated map.

Table 3. The precision-recall results of proposed algorithm in terms of object detected in maps.

Sequence	01	07	09	11	14
TP	71	47	68	35	8
FP	20	12	12	7	1
FN	10	1	2	5	4
Precision	78.0	79.7	85.0	83.3	88.9
Recall	87.7	97.9	97.1	87.5	66.7

Similarly, one can calculate the value of precision and recall for the proposed semantic map algorithm, except that data associations are determined by IoU in virtual observation space instead of in world coordinates. The results are shown in Table 3 ,which is based on IOU_{3D} not less than 35%. As we can see, the value of precision is no less than 78%, and the recall is larger than 66%, which shows the good performance of our algorithms.

5.3 Absolute object distance

One key evaluation aspect of 3D mapping methods is to know the positional error of 3D environments. For 3D object-centered semantic mapping methods, we propose to evaluate absolute object distance (AOD) between. The AOD metric calculates as follows. Given an estimated semantic map M^e, a ground-truth semantic map M^{GT}, and data association A that indicates the matched objects between the two maps, the metric calculates the average distance between two sets of matches objects. There are two variants of the metric. One calculates the distance directly, denoted as direct AOD (DAOD), which is useful for cases when absolute position error is important. Another one, denoted as aligned AOD (AAOD), first transfer estimated map M^e to be best aligned with the ground-truth map then calculate distance between them. The AAOD is more useful for cases where positioning between mapped objects is more important than the absolute positioning of a single object in a map.

$$\mathrm{DAOD}(M^e, M^{GT}, A) = \frac{\sum_{a\in A} ||M_a^e - M_a^{GT}||_2}{|A|}, \quad (10)$$

where $||\cdot||_2$ is the L^2 norm of a vector and M_a means taking center of an object from M with index in data association a.

$$\mathrm{AAOD}(M^e, M^{GT}, A) = \frac{\sum_{a\in A} ||T^A \cdot M_a^e - M_a^{GT}||_2}{|A|}, \quad (11)$$

where $T^A \in SE(3)$ is a 3D transformation that aligns M^e with M^{GT} by

$$T^A = \arg\min_{T^A} \sum_{a\in A} ||T^A \cdot M_a^e - M_a^{GT}||_2. \quad (12)$$

For the data association A, it can be calculated by matching IoUs between virtual observations on both maps when both the estimated camera trajectory and the ground-truth trajectory are available. In the worst case, it can be generated with human labeling.

5.4 Experiments

When evaluating absolute object distance metrics, data association between ground-truth maps and estimated maps is done through IoU score in virtual observations as stated in PRVO metric with IoU threshold of 60%. Results for the two AOD metrics are in Table 4.

Table 4. DAOD and AAOD of the proposed method

Sequence	DAOD (m)	AAOD (m)
01	3.55	1.18
07	6.37	2.42
09	17.06	7.59
11	1.71	0.56
14	0.40	0.12

Table 5. PRVO of the proposed method

Sequence	PRVO	
	Precision (%)	Recall (%)
01	50.0	59.9
07	36.6	47.4
09	44.1	48.7
11	58.5	62.6
14	61.1	74.0

5.5 Precision and recall of virtual observations

To evaluate the performance of object detection fusion performance while isolating the influence of camera trajectory estimation error from SLAM methods, we propose another metric called precision and recall of virtual observations (PRVO). Inputs for the PRVO metric are: an estimated semantic map M^e with corresponding estimated camera trajectory $\{T\}^e$ and ground-truth semantic map M^{GT} with ground-truth camera trajectory $\{T\}^{GT}$.

For each data frame, virtual observations are generated for both the estimated map and the ground-truth map by applying sensor model $s(\cdot)$

$$\begin{aligned} \{B\}_v^e &= s(M^e, T^e) \text{ for } T^e \in \{T\}^e \\ \{B\}_v^{GT} &= s(M^{GT}, T^{GT}) \text{ for } T^{GT} \in \{T\}^{GT}. \end{aligned} \quad (13)$$

An IoU threshold is used to determine the number of false positive (fp), false negative (fn), and true positive (tn) between $\{B\}_v^e$ and $\{B\}_v^{GT}$ for a frame. For all the frames, aggregated total numbers are calculated as $TN = \sum tn$, $FN = \sum fn$, and $TP = \sum tp$. Finally, a pair of precision and recall values are computed from TN, FN, and TP.

Experiments When evaluating PRVO metric, the IoU threshold is set to 50%. Results are reported in Table 5.

6. CONCLUSION

In this work, a simple and effective real-time 3D semantic mapping method is proposed. The proposed method takes per-frame bounding box detections and sensor (camera) extrinsic transformation estimates as inputs and produces a set of static 3D bounding boxes in a world coordinate system as 3D semantic mapping results.

To evaluate the proposed method, a new benchmark is derived from KITTI object tracking evaluation. Ground-truth semantic maps are constructed based on GPS/IMU data of KITTI and the labeled 3D bounding box. By fusing multiple annotation bounding boxes of the same object from frames, we get one single 3D bounding box for each object in the world frame as ground-truth semantic

maps. Three novel semantic map-centered metrics: DAOD, AAOD, and PRVO are also proposed. Experiments are conducted to evaluate the proposed method. The set of proposed method, metric and benchmarking dataset will serve as a new benchmark platform for easier comparison of new methods.

7. FUTURE WORK

Improvements can be made on the proposed baseline method. A major one is to include object appearance into the data association step so that it will be more robust for cluttered environments. There are other potential improvements, such as reducing the influence of false positive detections out of object detection methods, and adding more information, object property etc., to map objects.

In addition to the improvements on the proposed method itself. There are interesting work that can be done to feed information back to supporting object detections and camera trajectory estimation methods so that they can perform better with additional information made available.

REFERENCES

Bavle, H., De La Puente, P., How, J.P., and Campoy, P. (2020). VPS-SLAM: Visual Planar Semantic SLAM for Aerial Robotic Systems. *IEEE Access*, 8, 60704–60718. doi:10.1109/ACCESS.2020.2983121.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11618–11628. doi: 10.1109/CVPR42600.2020.01164.

Grinvald, M., Furrer, F., Novkovic, T., Chung, J.J., Cadena, C., Siegwart, R., and Nieto, J. (2019). Volumetric instance-aware semantic mapping and 3D object discovery. *IEEE Robotics and Automation Letters*, 4(3), 3037–3044. doi:10.1109/lra.2019.2923960.

McCormac, J., Handa, A., Davison, A., and Leutenegger, S. (2017). SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *Proc. of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. doi: 10.1109/icra.2017.7989538.

Nakajima, Y., Tateno, K., Tombari, F., and Saito, H. (2018). Fast and Accurate Semantic Mapping through Geometric-based Incremental Segmentation. In *Proc. of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. doi: 10.1109/iros.2018.8593993.

Pargi, M.K., Setiawan, B., and Kazama, Y. (2019). Classification of different vehicles in traffic using RGB and Depth images: A Fast RCNN Approach. In *Proc. of the 2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, 1–6. doi: 10.1109/IST48021.2019.9010357.

Qin, T., Chen, T., Chen, Y., and Su, Q. (2020). AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot. In *Proc. of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. doi: 10.1109/iros45743.2020.9340939.

Shi, S., Wang, X., and Li, H. (2019). PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–779. doi:10.1109/CVPR.2019.00086.

Weng, X., Wang, J., Held, D., and Kitani, K. (2020). 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. In *Proc. of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10359–10366.

Zhou, Z., Wang, M., Chen, X., Liang, W., and Zhang, J. (2019). Box Detection and Positioning based on Mask R-CNN for Container Unloading. In *Proc. of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, 171–174. doi: 10.1109/IAEAC47372.2019.8997535.