

Using Detection, Tracking and Prediction in Visual SLAM to Achieve Real-time Semantic Mapping of Dynamic Scenarios

Xingyu Chen¹, Jianru Xue^{1,†}, Jianwu Fang^{1,2}, Yuxin Pan¹ and Nanning Zheng¹

Abstract—In this paper, we propose a lightweight system, RDS-SLAM, based on ORB-SLAM2, which can accurately estimate poses and build semantic maps at object level for dynamic scenarios in real time using only one commonly used Intel Core i7 CPU. In RDS-SLAM, three major improvements, as well as major architectural modifications, are proposed to overcome the limitations of ORB-SLAM2. Firstly, it adopts a lightweight object detection neural network in key frames. Secondly, an efficient tracking and prediction mechanism is embedded into the system to remove the feature points belonging to movable objects in all incoming frames. Thirdly, a semantic octree map is built by probabilistic fusion of detection and tracking results, which enables a robot to maintain a semantic description at object level for potential interactions in dynamic scenarios. We evaluate RDS-SLAM in TUM RGB-D dataset, and experimental results show that RDS-SLAM can run with 30.3 ms per frame in dynamic scenarios using only an Intel Core i7 CPU, and achieves comparable accuracy compared with the state-of-the-art SLAM systems which heavily rely on both Intel Core i7 CPUs and powerful GPUs.

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) [1] is an important technique of perception and navigation for intelligent mobile systems, such as robots and autonomous vehicles. Due to the low cost, high resolution, and rich color information of camera, visual SLAM (vSLAM) has become an important research topic over the last years. Some excellent vSLAM systems have been established, such as ORB-SLAM2 [2], ElasticFusion [3], RTAB-Map [4].

However, classical vSLAM systems commonly assume that scenes are rigid and static, and this assumption leads to frequent failures of vSLAM systems in dynamic scenarios, where there are movable objects, such as people and cars. Even ORB-SLAM2 [2], one of the state-of-the-art vSLAM systems, may frequently fail in dynamic scenarios, and can only provide a map with incomplete descriptions. Its localization accuracy is also dramatically degraded. Obviously, these limitations are caused by movable objects in dynamic scenarios.

To overcome the effects of movable objects in dynamic scenarios to vSLAM systems, we propose three major improvements for ORB-SLAM2, and implement a robust and real-time vSLAM framework, RDS-SLAM, for mapping dynamic scenarios. The proposed RDS-SLAM can effectively

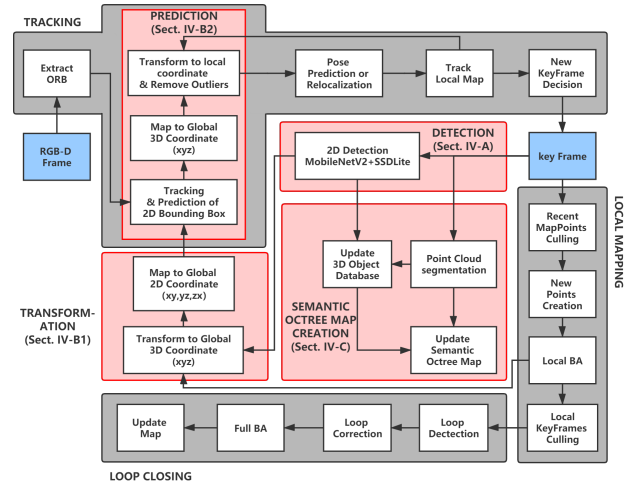


Fig. 1. The framework of RDS-SLAM. The threads filled with gray color are original threads of ORB-SLAM2. We denote the improvements over ORB-SLAM2 with red color. The threads in red color are the parallel improvements proposed in this paper. Additionally, a *Prediction* module is inserted into the *Tracking* thread of ORB-SLAM2.

remove the feature points belonging to movable objects, and build a semantic octree map at object level for complete description of dynamic scenarios.

More specifically, the proposed improvements, as well as major architectural modifications, are illustrated in Fig. 1. Firstly, we adopt a 2D object detection network as a parallel thread, which is denoted as *Detection* in Fig. 1, and the technical details are presented in Sect. IV-A. Instead of detecting in all frames as other dynamic SLAM systems do, we run it only in key frames to get the 2D movable objects.

Secondly, we propose an efficient prediction mechanism, which is denoted as *Transformation* and *Prediction* in Fig. 1. We transform the local 2D bounding box to global 3D coordinate and extend the classic local 2D tracking algorithm SORT [5] to global 3D coordinate to track 3D movable objects in key frames, and the constant velocity model is taken to predict other frames (Sect. IV-B). The running time of each frame of the prediction mechanism that we test on Intel i7 CPU is only 5ms.

Finally, we build *Semantic Octree Map Creation* as a parallel thread shown in Fig. 1 for both removing dynamic objects and creating a complete semantic map at object level. Instead of raising probability threshold of octree map like other state-of-the-arts do in dynamic scenarios, we use semantic information to distinguish whether the point clouds

*This work is partially supported by NSFC Projects 61751308 and U1713217.

¹The authors are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, P.R. China.

²The author is with the School of Electronic and Control Engineering, Chang'an University, Xi'an, P.R. China.

[†]Corresponding author's email: jrxue@mail.xjtu.edu.cn

are movable or not, then insert octree maps with different probabilities to remove the movable object (Sect. IV-C).

The rest of the paper is structured as follows: Section II discusses the related works. Section III presents an overview of RDS-SLAM. Three major improvements are detailed in Section IV, which are followed by experimental results in Section V. Finally, the paper is concluded with discussions and lines for the future in Section VI.

II. RELATED WORKS

There are many excellent vSLAM systems in literature for mapping scenarios [2], [3], [4] by using RGB-D data, and a comprehensive survey can be found in [1]. However, they often fail in dynamic scenarios, and this leads to many research works in recent years. In this section, we present a brief survey for these research efforts.

The core idea of improving vSLAM systems is to distinguish the dynamic parts of scenarios. For this purpose, it is straightforward to introduce segmentation [6], [7], [8], [9]. McCormac et al. [6] estimated poses and created a dense map through ElasticFusion [3], then built a single-frame map through a convolutional neural network (CNN) and finally merged two maps to generate a dense semantic map with higher classification accuracy than single-frame CNN. However, it cannot handle the dynamic scenarios. StaticFusion [7], Co-Fusion [8], and MaskFusion [9] had been proposed to deal with the dynamic scenarios. They focused on using segmentation information to directly build an accurate dense map that can distinguish the dynamic objects and static scenarios. However, these works have relatively low localization accuracy and heavily rely on intensive computation efficiency.

Among many vSLAM works, ORB-SLAM2 [2] is widely accepted as the best open source vSLAM system with high localization accuracy and map reusability, but it also fails in dynamic scenarios. The situation has been significantly improved by DynaSLAM [10] and DS-SLAM [11], which are two important variants of ORB-SLAM2. To remove the ORB [12] feature points of dynamic objects, DynaSLAM serially added *Mask R-CNN* [13], *Low-Cost Tracking* and *Multi-view Geometry* to the front of ORB-SLAM2 before extracting the ORB feature points. However, since it serially added three modules to the front of ORB-SLAM2, the average time it took per frame using CPU+GPU is about 500 *ms*. Similar to DynaSLAM, DS-SLAM also serially added *Moving Consistency Check* module and *Remove Outliers* module to the *Tracking* thread of ORB-SLAM2. Different from DynaSLAM, DS-SLAM parallel added *SegNet* [14] thread and *Dense Map Creation* thread to ORB-SLAM2. It finally combines the results of parallel *SegNet* thread and the serial *Moving Consistency Check* module in each frame. Even with such a parallel architecture, its average time of processing a frame using CPU+GPU is about 59.4 *ms*. In summary, neither DynaSLAM [10] nor DS-SLAM [11] can work in real time without GPUs, and thus cannot meet with lightweight applications.

Motivated by the aforementioned works, we propose the real time RDS-SLAM, which can build a complete semantic octree map of dynamic scenario without using GPUs as well as the competitive accuracy compared with DynaSLAM [10] and DS-SLAM [11].

III. SYSTEM OVERVIEW

We propose a real-time and lightweight RGB-D vSLAM system in dynamic scenarios based on ORB-SLAM2 [2]. We use object detection and object tracking only in key frames, and use low-cost prediction in other frames to reduce the computational cost, as shown in Fig. 1.

In addition to *Tracking*, *Local Mapping* and *Loop Closing*, three parallel threads of original ORB-SLAM2, we add *Detection*, *Transformation* and *Semantic Octree Map Creation*, three parallel threads into the system. And we also insert a new module named *Prediction* into the *Tracking* thread.

After the processing of *Extract ORB*, *Pose Prediction or Relocalization* and *Track Local Map* in *tracking* thread, ORB-SLAM2 has realized a visual odometry that can estimate the pose transformation between frames in a static scenario. In order to build the map and optimize the pose, ORB-SLAM2 proposes *New KeyFrame Decision* module to select key frames from visual sequence and put them into *Local Mapping* thread. The mechanism of *New KeyFrame Decision* emphasizes that when the scenario changes, a key frame will be inserted after a certain time interval is met, and when the scenario changes quickly, a key frame will be directly inserted regardless of the time interval.

In RDS-SLAM, we believe that detecting objects in all frames without a selection will cost too many computational resources, because the scenario does not always change during localization and mapping of robots. In other words, we should use *Detection* thread only when scenario changes and increase the frequency of detection when scenario changes quickly. Thus we can utilize the mechanism of *New KeyFrame Decision* as the mechanism of *Detection* to realize the adaptive computational resource allocation of *Detection* thread by using *Detection* only in key frames instead of all the frames.

After *New KeyFrame Decision* putting the key frames into the *Local Mapping* thread, ORB-SLAM2 will check the recently added feature points on the map (map points) by *Recent MapPoints Culling* module, as shown in Fig. 1. It emphasizes that if a map point is constructed, it must be observed by the next three key frames. ORB-SLAM2 effectively eliminates the incorrect map points through *Recent MapPoints Culling*, but it cannot effectively remove the map points on movable objects.

On the contrary, in RDS-SLAM, the map points on movable objects of the latest key frame were temporarily built into the map. After RDS-SLAM detecting the latest key frame using an object detection network and putting the results into each of the future frames, the map points on movable objects built by the latest key frame will no longer be observed in the next key frame. Thus RDS-SLAM

Algorithm 1 Algorithm of 3D object prediction

Input: The bounding box positions matrix, $P_w(x, y, z)$; or *empty*;
Output: Prediction of bounding box positions matrix, $P_{wp}(x_p, y_p, z_p)$;
1: Mapping matrix P_w to global 2D coordinate planes (xOy, yOz, zOx), as $P_{xy}(x, y)$, $P_{yz}(y, z)$ and $P_{zx}(x, z)$;
2: **if** *input* \neq *empty* **then**
3: Using SORT which includes constant velocity Kalman filter framework and Hungarian algorithm to track and predict the positions of P_{xy} , P_{yz} , P_{zx} simultaneously and update them;
4: **else**
5: Using constant velocity model to predict positions of P_{xy} , P_{yz} , P_{zx} in the current frame and update them;
6: **end if**
7: Find the max matrix of P_{xy} , P_{yz} , P_{zx} , name it as $P_{ab}(a, b)$, and $P_{bc}(b, c)$, $P_{ca}(a, c)$ for others;
8: Use the information of latest frame to generate $P_{abw}(a, b, 0)$, $P_{bcw}(a_l, b, c)$ and $P_{caw}(a, b_l, c)$;
9: **for** each p_{abw} in P_{abw} **do**
10: **for** each p_{bcw} , p_{caw} in P_{bcw} , P_{caw} **do**
11: Calculate IOU of p_{abw} with p_{bcw} , p_{caw} in coordinate aOb ;
12: Record the largest IOU and its p_{index} in p_{bcw} or p_{caw} ;
13: **end for**
14: The c of $p_{abw}(a, b, c)$ is equal to the c of p_{index} ;
15: **end for**
16: $P_{wp}(x_p, y_p, z_p) = P_{abw}(a, b, c)$;

C. Semantic Octree Map Creation

To provide a complete map description, we build a semantic octree map to show whether a location is occupied, what an occupied point represents for in object level and remove the movable occupied point at the same time. Firstly, we use *Point Cloud Segmentation* module to segment and cluster 3D point cloud to obtain 3D object's boundary. Then we use *Update 3D Object Database* module to project it to the 2D bounding box to create the 3D bounding box. The *Point Cloud Segmentation* and *Update 3D Object Database* module comes from Ewenwan's open source code [20] which is based on Point Cloud Library (PCL) [21] to slowly build a 3D object database from 3D point cloud and 2D bounding box, so we create the semantic octree map in key frames.

The *Update Semantic Octree Map* module proposed by us will firstly filter the current point cloud and put it into the candidate occupied point cloud. After that, the module will query whether a candidate occupied point is located in the 3D bounding box. If it is true, it will be assigned the corresponding color according to its label. Otherwise, it will retain the original color. In the meanwhile, the module will query whether the candidate occupied point is located in the 3D bounding box of movable objects. If it is, it will be marked as a movable point.

Last but the most important part is updating the probability of occupation of octree map. The octree map [22] uses probability to indicate whether a node is occupied, and its occupation probability will be updated when there is a new observation. Assuming n is one of these nodes, then we map the occupancy probability $P(n) \in [0, 1]$ of the node n to the logistic regression variable $L(n) \in \mathbb{R}$ and then the update is performed on the space \mathbb{R} . The mapping can be calculated as $L(n) = \log\left(\frac{P(n)}{1-P(n)}\right)$. Assume that the observation at time T is Z_T , then the updated formula for occupation probability of the node n can be mapped to:

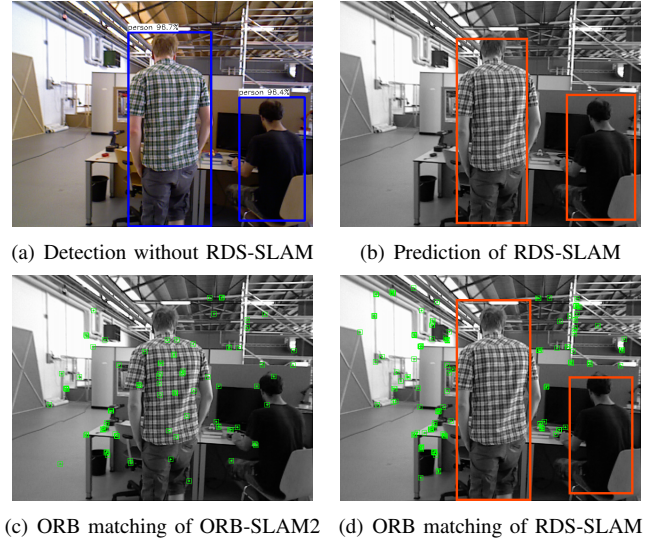


Fig. 4. Test results on w_{xyz} sequence. Fig. 4(a) shows the test results of the current frame using detection alone without RDS-SLAM. Fig. 4(b) shows the result of the current frame using *Prediction* of RDS-SLAM. We could see that the result of *Prediction* is very close to the result of detecting alone. Fig. 4(c) shows the result of the ORB matching of ORB-SLAM2. Fig. 4(d) shows the result of ORB matching of RDS-SLAM. It is obvious that the result of ORB matching of RDS-SLAM has effectively removed the feature points on the movable objects and got more ORB matching of static scenario.

$$L(n|z_{1:T}) = L(n|z_{1:T-1}) + L(n|z_T). \quad (2)$$

In general, if a node n is inserted by any point at time T , $L(n|z_T) = \tau$, and if not, $L(n|z_T) = 0$. The occupation probability threshold is p , while the principle that the node n is considered to be occupied is $L(n|z_{1:T}) > \log\left(\frac{p}{1-p}\right)$.

Previous SLAM systems with octree map remove moving objects by increasing threshold p . These methods can create an octree map without moving objects, but cannot remove potentially movable objects. And after increasing threshold p , when the robot moves fast and the observation time for each scenario is limited, the static scenarios that are less observed will also be removed.

To this end, RDS-SLAM removes the movable objects by inserting the points marked as movable or static with different probability. So that it could remove all the movable objects including moving objects and potentially movable objects, and it will not remove the less observed static scenarios at the same time.

In our experiment, we set the occupancy probability threshold to the default $p = 0.5$. For the points marked as movable, we set $L(n|Z_T) = -0.41$, for other points, we set $L(n|Z_T) = 0.85$, if unobserved, $L(n|Z_T) = 0$.

V. EXPERIMENTAL RESULTS

A. Dataset, Experimental Setting and Evaluation Metrics

We implement RDS-SLAM on a Ubuntu operation system running on an Intel i7 CPU without any GPU accelerators. TUM RGB-D dataset [23] is used for evaluation and comparison of RDS-SLAM with other state-of-the-art vSLAM

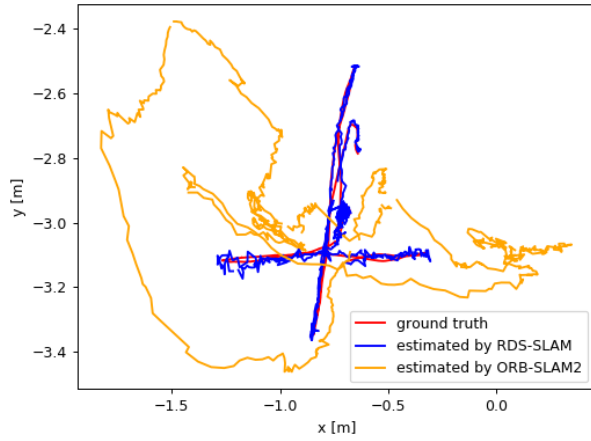


Fig. 5. Test results on w_{xyz} sequence. Figure shows the estimated trajectories of ORB-SLAM2 and RDS-SLAM, which are denoted with colors of orange and blue respectively, while the ground truth marked by red color is given.

systems. The dataset contains four high-dynamic sequences w_{xyz} , w_{static} , w_{rpy} and $w_{halfsphere}$ which have two walking men in front of the cameras. It also has two low-dynamic sequences s_{xyz} and s_{static} which have two sitting men in front of the cameras. The root mean square error ($RMSE$) of the estimated trajectory with respect to the ground truth is used as the accuracy metric of a vSLAM system [23]. To avoid the impact of non-deterministic nature, we calculate the $RMSE$ of RDS-SLAM with six dynamic sequences, and run the system with each sequence for 10 times. Then we use the median $RMSE$ of each sequence as the accuracy metric of RDS-SLAM.

Additionally, we also compare the efficiency of RDS-SLAM with the state-of-the-arts in term of computational cost, which is measured with processing time per frame by taking the computing platform into consideration.

B. Analysis and Discussions

For comparison of RDS-SLAM with ORB-SLAM2, we test both of them using the sequence w_{xyz} . Fig. 5 shows both the estimated trajectories of ORB-SLAM2 and RDS-SLAM as well as the ground truth. It clearly shows that the estimated trajectory of RDS-SLAM coincided well with the ground truth, while ORB-SLAM2 fails most of the time.

We then compare RDS-SLAM with 7 vSLAM systems [2], [3], [7], [8], [9], [10], [11] where StaticFusion [7], Co-Fusion [8], MaskFusion [9], DynaSLAM [10] and DS-SLAM [11] are the state-of-the-art dynamic SLAM systems. Table I shows experimental results in terms of $RMSE$. Results of RDS-SLAM come from our experiments, others are from reports in [7], [9], [10], [11]. The two best results are shown in bold in Table I. It shows that the accuracy of DynaSLAM, DS-SLAM and RDS-SLAM is higher than other SLAM systems.

We also compare the efficiency of RDS-SLAM with DynaSLAM and DS-SLAM since they have comparable localization accuracy and have higher accuracy than other systems relying on powerful GPUs. As shown in Table I,

both DynaSLAM and DS-SLAM need CPUs accelerated with GPUs to achieve the speed of 500 ms and 59.4 ms per frame, while RDS-SLAM can achieve a speed of 30.3 ms per frame only with an Intel i7 CPU.

Furthermore, we analyze the computational cost of the top 3 accurate vSLAM systems in details, since they all are improved from ORB-SLAM2. As shown in Table II, where SITC denotes the serial increase in time consumption, PISC specifies the parallel increase in time consumption, and ATCPF is the average time consumption per frame. SITC is the key point to determining whether an improved system can run in real time. In DynaSLAM, it serially added *Low-Cost Tracking*, *Multi-view Geometry* and *Mask R-CNN* to improve ORB-SLAM2, which causes an increment of 500 ms of ATCPF. In DS-SLAM, it introduces *SegNet* in parallel, but each current frame needed to wait for the result of *SegNet*. In the meanwhile, DS-SLAM serially added the *Moving Consistency Check* into ORB-SLAM2. They cause an increment of 30ms of ATCPF. Different from detecting objects in all frames, RDS-SLAM uses the detection results of key frames to predict the positions of movable objects in other frames. Table II shows that RDS-SLAM adds MobileNetV2 SSDLite in parallel, and each current frame uses the output from the lightweight *Prediction* without waiting, so the ATCPF of RDS-SLAM is only increased by about 5ms.

Fig. 6 shows the semantic octree maps built by RDS-SLAM. Firstly, we test it in w_{xyz} sequence. Fig. 6(a) shows the test result of the octree map without semantic association. Fig. 6(b) demonstrates the result of the octree map with semantic association. It is obvious that the octree map with semantic association effectively removes movable objects. Fig. 6(c) shows the semantic octree map we build for the *room* sequence comes from TUM RGB-D dataset [23], and it demonstrates that both the geometric information and semantic information are completely presented in the map.

VI. CONCLUSIONS

In this paper, we present an efficient RDS-SLAM system, which is a lightweight visual semantic SLAM system for dynamic scenarios. It runs well with an Intel i7 CPU in real time without using any GPU accelerator. RDS-SLAM is an improved variant of ORB-SLAM2 by the adoption of *Detection*, *Transformation*, *Prediction* and *Semantic Octree Map Creation*, which efficiently removes the feature points belonging to movable objects and builds an accurate semantic octree map at object level for dynamic scenarios.

The efficiency of RDS-SLAM was validated with the TUM RGB-D dataset. We also compare it with the state-of-the-art vSLAM systems. Experimental results show that RDS-SLAM can run with 30.3 ms per frame using only an Intel i7 CPU and reach the competitive performance to the state-of-the-art SLAM systems in dynamic scenarios.

Future extensions of this work might be using semantic description to clarify ambiguity in corresponding feature points, and exploring geometric structures to handle high dynamic scenarios.

TABLE I
COMPARISON OF $RMSE(cm)$ AND COMPUTATIONAL COST IN DYNAMIC SCENARIOS.

Sequence	ORB-SLAM2	ElasticFusion	Co-Fusion	StaticFusion	MaskFusion	DynaSLAM		DS-SLAM		Ours (RDS-SLAM)	
	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	Cost	RMSE	Cost	RMSE	Cost
w_xyz	45.9	90.6	69.6	12.7	10.4	1.5		2.5		1.7	
w_static	9.0	29.3	55.1	1.4	3.5	0.6		0.8		0.9	
w_rpy	66.2	-	-	-	-	3.5	≈ 500 ms	44.4	59.4 ms	3.9	30.3 ms
w_half	35.1	63.8	80.3	39.1	10.6	2.5	on CPU +	3.0	on i7 CPU+	3.1	on i7 CPU
s_static	0.9	0.8	1.1	1.3	2.1	-	M40 GPU	0.7	P4000 GPU	0.8	
s_xyz	0.9	2.2	2.7	4.0	3.1	1.5		-		1.1	

TABLE II
COMPUTATIONAL COST ANALYSIS.

Framework	Platform	SITC	PITC	ATCPF
DynaSLAM	CPU + M40 GPU	LC Tracking 1.64 ms MV Geometry 285 ms Mask R-CNN 195 ms	0	≈ 500 ms
DS-SLAM	i7 CPU + P4000 GPU	MC Check 29.5 ms	SegNet 37.6ms	59.4 ms
RDS-SLAM	i7 CPU	Prediction ≈ 5 ms	MV2SSD 53 ms	30.3 ms
ORB-SLAM2	i7 CPU	0	0	25.6 ms

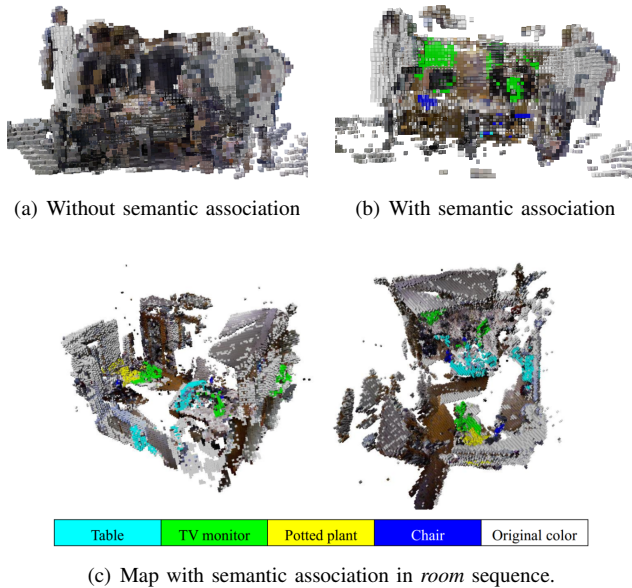


Fig. 6. Octree maps built by RDS-SLAM.

REFERENCES

- [1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.
- [2] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [3] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "Elasticfusion: Real-time dense slam and light source estimation," *The International Journal of Robotics Research*, vol. 35, no. 14, pp. 1697–1716, 2016.
- [4] M. Labbé and F. Michaud, "Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.
- [5] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3464–3468.
- [6] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4628–4635.
- [7] R. Scona, M. Jaimez, Y. R. Petillot, M. Fallon, and D. Cremers, "Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–9.
- [8] M. Rünz and L. Agapito, "Co-fusion: Real-time segmentation, tracking and fusion of multiple objects," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4471–4478.
- [9] M. Rünz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2018, pp. 10–20.
- [10] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4076–4083, 2018.
- [11] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Dslam: A semantic visual slam towards dynamic environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1168–1174.
- [12] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, vol. 11, no. 1. Citeseer, 2011, p. 2.
- [13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.
- [14] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [16] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7310–7311.
- [17] Tencent, "Ncnn," <https://github.com/Tencent/ncnn>, last accessed 10 May 2019.
- [18] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
- [19] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [20] Ewenwan, "Homepage," <https://github.com/Ewenwan>, last accessed 10 May 2019.
- [21] R. B. Rusu and S. Cousins, "Point cloud library (pcl)," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1–4.
- [22] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.
- [23] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 573–580.