EasyChair Preprint

№ 4149

Visual SLAM Technology Based on Weakly Supervised Semantic Segmentation in Dynamic Environment

Jianxin Liu, Menglan Zeng, Yuchao Wang and Wei Liu

September 7, 2020

Visual SLAM technology based on weakly supervised semantic segmentation in dynamic environment

Jianxin Liu, Menglan Zeng, Yuchao Wang, Wei Liu
(School of Mechanical Engineering, Xihua University, Chendu 610039, China)

ABSTRACT

A visual simultaneous localization and mapping system in a dynamic environment is usually vulnerable to wrong associated data caused by the moving targets, which causes a large error in the pose estimation of the mobile robot and affects the subsequent tasks of the robot. Combining semantic segmentation information to remove dynamic feature points in the image is an effective method to improve the accuracy of the SLAM system. However, the existing visual SLAM based on semantic segmentation usually adopts the fully supervised approaches to segment the dynamic scenes, which depends on a large number of training data sets with labelled information to guarantee accuracy and limits the application of SLAM system. To address this issue, a visual semantic SLAM system that applies weakly supervised semantic segmentation to dynamic scenes is proposed to broaden the application range of the system. Firstly, the system extracts the feature points of the input image and checks the moving consistency, and then segments the dynamic target with the weakly supervised methods. Secondly, the semantic segmentation results are used to remove the dynamic feature points in the image. Finally，the stable static feature points are adopted to carry out the pose estimation. Furthermore, this paper also uses the Automatic Color Equalization algorithm to pre-process the input image, which enhances the contrast of the image, and improves the accuracy of weakly supervised semantic segmentation. Experiments were performed on the public TUM data sets and real environment. The results show that the accuracy of the SLAM system based on the weakly supervised network adopted in this paper is significantly higher than that of the traditional ORB-SLAM2 system, and also higher than the SLAM system of the weakly supervised network DSRG. The accuracy is close to the fully supervised semantic SLAM system.

Keywords: dynamic environment, weakly supervision semantic segmentation, moving consistency check, semantic SLAM

1. INTRODUCTION

Simultaneous localization and mapping (SLAM) is an important part of robot autonomous navigation system. In the SLAM system, environmental information can be collected by means of sensors, such as Sonar, Lidar, Camera and IMU, which can help the robot to estimate the position and build the map of the surrounding environment [1]. Due to the low cost of visual sensors and the large amount of image information, the research of vision SLAM system has become one of the hot topics in the field of computer vision and robotics. Visual SLAM system usually consists of four parts: Visual Odometry, Optimization, Loop Closing and Mapping [2]. At present, many excellent algorithms have been proposed, such as ORB-SLAM2[7], LSD-SLAM [5] and SVO [6].

The current vision SLAM algorithm is mostly based on the assumption of static environment, but it is difficult to keep the

static assumption in the real environment. The assumption of static environment will lead to incorrect information association, and even lead the location failure of the mobile robot.

There are two solutions to the SLAM system in dynamic scenarios: using the algorithm of motion detection [9], [10], [11], or using the method of deep learning [12], [14], [16], [18]. Although motion detection algorithms are used to deal with the problem of dynamic environment, these algorithms are usually complex and limited in special applications. In recent years, with the successful application of deep learning in image recognition and image segmentation, the research of semantic segmentation in visual SLAM has gradually attracted attention. Furthermore, the robot can enhance the understanding of the surrounding environment through the segmented semantic information, and then remove the dynamic targets in the dynamic scene, which is helpful to improve the accuracy and robustness of the SLAM system. Therefore, semantic segmentation of moving objects is an effective method to improve the performance of dynamic visual SLAM system. However, the general semantic segmentation method is based on the fully supervised method requiring a large number of annotated images in the training process, which is time-consuming and expensive and limits the applicability and adaptability of the convolutional neural networks in visual SLAM problems.

In this paper, we propose to adopt the weakly supervised semantic segmentation method [22] to solve the time-consuming dynamic object detecting problem. The network is pre-trained from image-level label images, and Condition Random Field(CRF) is used to optimize the edge of segmentation results. In addition, we uses the ACE image pre-processing method to improve the seed quality of the network and further improve the accuracy of weakly supervised semantic segmentation. And then, weakly supervised semantic segmentation network is adopted to segment the dynamic targets in the scene and the result is passed to the tracking thread of ORB-SLAM2 system, so that the system can effectively avoid the impact of dynamic targets.

The highlight is that the ACE improved weakly supervised semantic segmentation method [22] is adopted to solve the dynamic scene problem of visual SLAM. Furthermore, the weakly supervised semantic segmentation network is integrated with the ORB-SLAM2 system to avoid the impact of dynamic targets. At the same time, the training cost of semantic segmentation is reduced, and the robustness and stability of visual SLAM system are improved. In this paper, the fully supervised semantic segmentation method and the weakly supervised semantic segmentation method are compared under the architecture of ORB-SLAM2 system, and the real-time performance and effectiveness of this method are verified on the TUM RGB-D dataset.

In the rest of this paper, the structure is as follows. Section 2 reviews previous related works. The weakly supervised semantic segmentation method used in our system is described in section 3. The section 4 introduces the system structure of this paper in detail, and combines the semantic segmentation method with the ORB-SLAM2 system. The experimental results in section 5 prove the effectiveness of the proposed method. Finally, a brief conclusion is given in section 6.

2. RELATED WORK

2.1 Visual SLAM

Visual SLAM has achieved rapid development in the last decade. It has drawn attention of researchers because of its low cost. Davisonet al. [3] proposed the first real-time SLAM system-Mono SLAM, which is based on extended Kalman filter. However, the filter-based method is prone to cumulative errors and is not suitable for large-scale scenes. Klein et al. [4] proposed a keyframe-based visual SLAM system--PTAM, which became the benchmarks for the follow-up SLAM

research because of proposing to simultaneously handle the tracking task and the mapping task on two threads. Engel et al. [5] proposed a monocular SLAM system LSD-SLAM based on the direct method, which is suitable for large-scale scenes and can create a semi-dense 3D environment map. Forster et al. [6] proposed a fast Semi-direct monocular Visual Odometry (SVO), which combines the advantage of feature point based method and direct tracking optical flow one. Mur-artal et al. improved the ORB-SLAM [7] system by adding binocular camera and RGB-D depth camera to the ORB-SLAM2 [8] system, which are successful applications of SLAM based on the feature point tracking method.

While modern SLAM system has been successfully demonstrated mostly in a static environment, unexpected changes of surroundings would probably corrupt the quality of the state estimation and even lead to system failure, such as walking people and moving vehicles.

2.2 Semantic SLAM for dynamic scene

The development of deep learning provides some solutions for visual SLAM system in dynamic scenes. YU et al. [12] proposed the DS-SLAM system by combining the SegNet [13] with ORB-SLAM2. The system operates on each image frame to segment the dynamic target and remove the dynamic feature points. Bescos et al. [14] used Mask-RCNN [15] to detect the dynamic target. Xi et al. [16] used PSPNet [17] to process dynamic scenes. Although the Mask-RCNN, Segnet [13] and PSPNet [17] have achieved high performance in dynamic scenes, they have to be trained under fully supervised method, which requires a large amount of time-consuming pixel-wise annotations for training. Xiao et al. [18] used SSD [19] detector with prior knowledge to detect dynamic targets. Sun et al. [27] improved weakly supervised semantic segmentation network SEC [20] with super-pixel to segment dynamic targets in scenes, but this method takes a long time and affects the real-time nature of the SLAM system.

2.3 Weakly supervised semantic segmentation

Weakly supervised semantic segmentation methods can reduce the annotation burden. According to the types of annotation required by the system, existing weakly-supervised methods are based on various annotations such as bounding box[28], point[29]，scribble[30] and image-level label[20],[21],[22]. Since the image-level labels are abundant and relatively cheap to collect, we train a semantic segmentation network[22] starting from some high-quality seeds, and a more accurate segmentation result is obtained.

3. ADOPTED WEAKLY SUPERVISED SEMANTICSEGMENTATION

This paper uses a semantic segmentation network[22] trained with image-level label. The model uses the Saliency Guided Self-attention Network (SGAN) to generate dense and accurate seeds. And utilize the high-quality seeds as ground-truth to train the weakly semantic segmentation network.

The overview of our adopted segmentation system is shown in Figure 1, Firstly, this system adopts Automatic Color Equalization (ACE) method to preprocess the training image to improve the contrast of the image. Then the processed image are fed into the SGAN network module [22] to obtain high-quality image seeds. Finally, high-quality image seeds are used to train the Deep Seeded Region Growing (DSRG) [21] segmentation network.

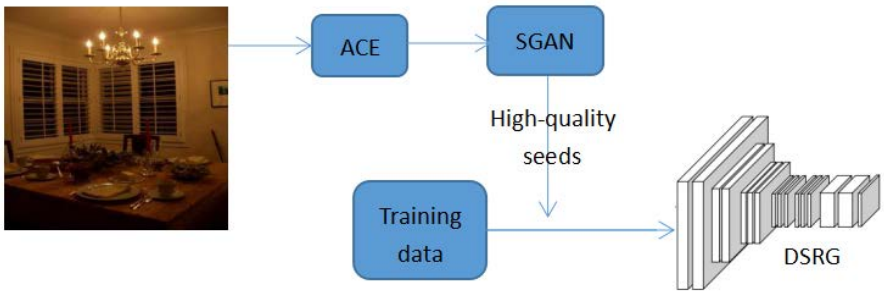


Figure 1. The overview of the adopted weakly supervised semantic segmentation system.

The initial seeds of DSRG network [21] are sparse and cannot provide information for segmentation. Therefore，SGAN network [22] is used to improve the quality of seeds and improve the semantic segmentation accuracy. In addition, ACE is used in this paper to pre-process the images, which improve the image contrast, make seeds easier to be detected and further improve the semantic segmentation accuracy. Figure 2 shows the segmentation seeds of DSRG [21] and our method. As what we can see, the segmentation seeds of DSRG are sparse and cannot provide rich pixel information. But our method produces a dense and accurate segmentation seeds.

(a) Image (b) DSRG (c) Our

Figure 2. The example of segmentation seeds on training data

The weakly supervised semantic segmentation network of this paper has achieved 63.54% mIoU on the validation set of

PASCAL VOC 2012. Compared with the segmentation results of the weakly supervised segmentation network DSRG [21] , the model in this paper improves the accuracy of semantic segmentation by about 5.11%.

4. SYSTEM FRAMEWORK

The open source ORB-SLAM2 is a relatively complete system in visual SLAM, which is easy to combine with other functions and is widely used. Because the ORB-SLAM2 system adopts the optimization module to reduce the drift and cumulative error in the trip, the system has good stability in the static environment. However, the ORB-SLAM2 system often mis-correlate the feature points on the dynamic target in the dynamic environment, which affect the stability of the system. Aiming at this problem, this paper adds a weakly supervised semantic segmentation model on the framework of ORB-SALM2. Firstly, the weakly supervised semantic segmentation network is used to detect and segment dynamic targets, and then the segmentation results are used to remove the dynamic feature points in the scene to optimize the visual odometer module. Finally, a visual semantic SLAM system for dynamic scenes is obtained. The system architecture is shown in Figure 3.

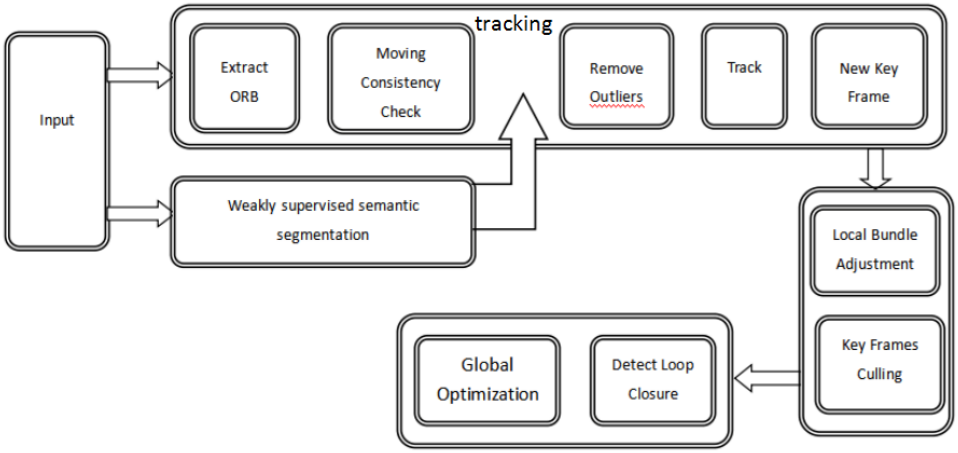


Figure 3. The framework of the proposed system.

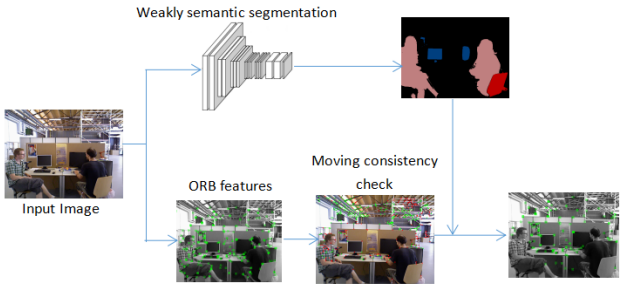


Figure 4. The elimination process of dynamic feature points.

Figure 4 shows the elimination process of dynamic feature points. The system passes the image frame into the tracking

thread and semantic segmentation thread at the same time. In the tracking thread, ORB features are extracted from the image frame and the moving consistency is checked. In the semantic segmentation thread, the weakly supervised network model is used to detect and segment the dynamic targets, and the segmentation results are passed into the tracking thread. Then the weakly supervised semantic segmentation results were combined with the moving consistency checking results to eliminate the feature points on the dynamic targets in the environment, and then the reliable feature points were matched and the pose estimation was performed. The neural network optimized SLAM system avoids the influence of the feature points on the dynamic targets and improves the accuracy and efficiency of camera pose estimation.

5. EXPERIMENTS

5.1 Experiment setup

Our method is implemented based on ORB-SLAM2 [7] with default parameters, and this method is tested on the Dynamic Objects sequences of TUM RGB-D dataset [26]. The experiments in this article are all done with Intel i7-8700 CPU and a GeForce GTX 1660Ti GPU card. Our segmentation network is implemented in Caffe.

In this paper, we take the human as a typical representative of dynamic objects. The sequences of TUM RGB-D dataset contain RGB images and depth images, and also provide the ground real trajectory. By comparing the trajectory obtained in the experiment with the real trajectory on the ground, the error of the SLAM system can be determined effectively. For the Dynamic objects sequences of TUM RGB-D dataset [26] in this article, the xyz represents rgbd_dataset_ freiburg3_walking_xyz, the sitting represents rgbd_dataset_freiburg3_sitting_static and the halfsphere represents rgbd_ dataset_reiburg3_walking_halfsphere. Among these TUM sequences, the sitting sequences depict low-dynamic scenes, and the other sequences depict high-dynamic scenes.

This paper includes three groups of tests: 1) Comparing the effect of removing dynamic feature points of images with different semantic segmentation networks, we verify the effectiveness of the weakly supervised segmentation network in this paper. 2) The absolute trajectory error and relative pose error of the semantic SLAM system in this paper were compared with that of the typical SLAM system to verify that the proposed method can improve the robustness of the SLAM system. 3) The real-time performance of semantic SLAM system in this paper was verified by the time of single frame processing in different network models.

5.2 Removing dynamic feature points

We use the fully supervised segmentation network SegNet[13], the weakly supervised segmentation network DSRG[21] and the weakly supervised segmentation network in this paper[22] to process the dynamic feature points of image frames. Firstly, the traditional method is used to extract the ORB feature points in the image, and the moving consistency of the image is checked, then we combine the results with the semantic segmentation results, and remove the dynamic feature points. Finally, the error of the system's pose estimation is reduced, and the stability of the system is improved.

Some sample frames marked with detected feature points are shown in Figure 5. The first row shows the feature points detected by ORB-SLAM2[7], the second row shows the feature points detected by the ORB-SLAM2 based on SegNet, the third row shows the feature points detected by the ORB-SLAM2 based on DSRG, and the bottom row shows the feature points detected by the ORB-SLAM2 based on our network. Both the SLAM system using SegNet and our network can remove the feature points on the dynamic targets well, but the weakly supervised network model DSRG cannot completely segment the dynamic target, so the SLAM system of the DSRG network model cannot completely remove the dynamic

feature point. It can be seen that the weakly supervised segmentation network proposed in this paper is close to the level of fully supervised segmentation network in the task of removing dynamic feature points.

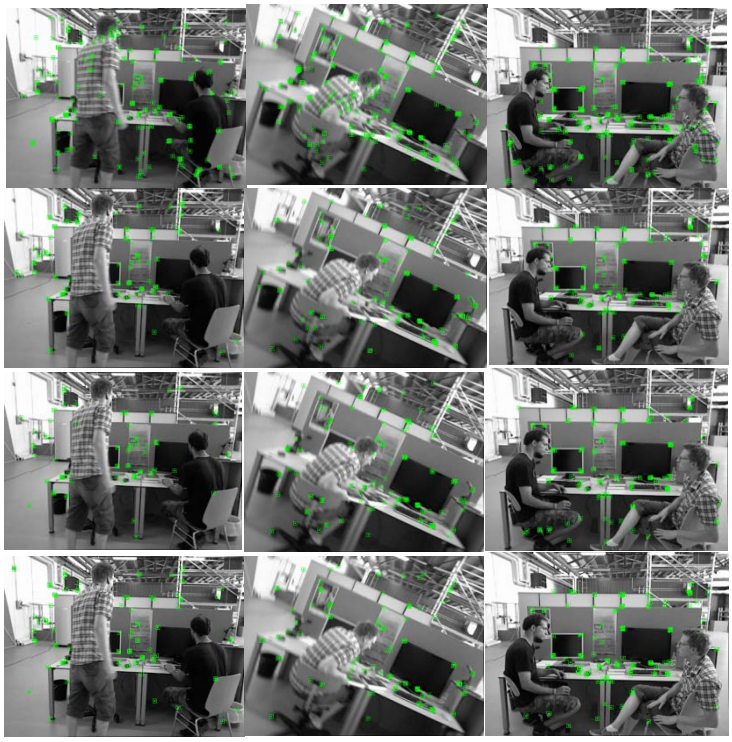


Figure 5. Sample frames marked with detected feature points. The first row shows the feature points detected by ORB-SLAM2, from the second row to the bottom row shows the feature points detected by SegNet[12], DSRG[21] and our network[22].

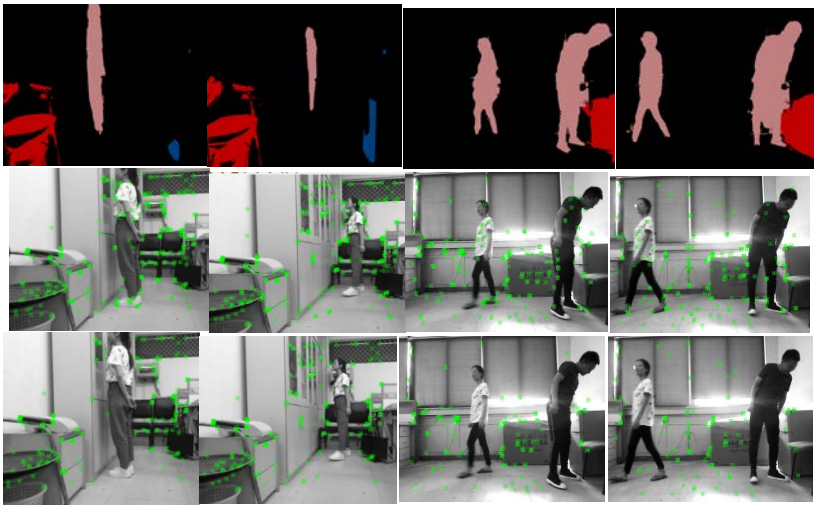


Figure 6. Experimental result in our lab environment

To demonstrate the robustness of our system, we conducted experiments in our laboratory environment. The images are captured by Kinect V2. Figure 6 shows the results of outlier's rejection. The sub-figures from top row to bottom row are semantic segmentation results, ORB feature extraction results, and images after outlier's removal respectively. The pink segmented region represents the dynamic target area. As we can see, the dynamic features points that fall in the dynamic

target area are removed, and the stable feature points are retained.

5.3 Evaluation using TUM RGB-D dataset

We employed the metrics Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) for the quantitative evaluations for the different SLAM systems. The ATE stands for global consistency of trajectory, and the RPE measures the translational drift error. We present the values of RMSE, Mean Error, Median Error and S.D. in this paper.

The quantitative comparison results are shown in Table I. It can be seen that the proposed method out-performs ORB-SLAM2. The performance of our proposed method is better than that of the weakly supervised network model DSRG, and close to that of the fully supervised network model SegNet. The results indicate our proposed method can improve the robustness and stability of SLAM system in high-dynamic environments significantly. However, in low-dynamic sequences, for example, the sitting sequence, the improvements of performance are not obvious. We think the reason is that ORB-SLAM2 can handle the low-dynamic scenes and achieve good performance, so the improvement of our method is limited.

TableII. Experimental result on dynamic sequences on TUM RGB-D dataset.

Squence	**Methods**	**Absolute Trajectory Error (ATE)**				**Translational Relative Pose Error (RPE)**			
		RMSE	**Mean**	**Median**	**S.D**	**RMSE**	**Mean**	**Median**	**S.D**
xyz	ORB-SLAM2	0.7209	0.6404	0.5880	0.3259	0.3952	0.2994	0.2399	0.2567
	SegNet	0.0313	0.0230	0.0190	0.0211	**0.0399**	**0.0265**	**0.0191**	0.0297
	DSRG	0.0512	0.0449	0.0424	0.0234	0.0425	0.0294	0.0212	0.0299
	Ours	**0.0205**	**0.0166**	**0.0141**	**0.0120**	0.0406	0.0317	0.0265	**0.0243**
Halfsphere	ORB-SLAM2	0.5401	0.4539	0.4235	0.2909	0.3837	0.2357	0.0771	0.3026
	SegNet	0.0308	0.0259	0.0219	0.0166	0.0310	0.0263	0.0245	0.0162
	DSRG	0.0354	0.0297	0.0254	0.0192	0.0357	0.0303	0.0266	0.0187
	Ours	**0.0298**	**0.0247**	**0.0207**	**0.0166**	0.0361	0.0311	0.0276	0.0183
Sitting	ORB-SLAM2	0.0082	0.0072	0.0065	0.0039	0.0093	0.0082	0.0073	0.0044
	SegNet	**0.0064**	**0.0055**	**0.0048**	**0.0033**	**0.0073**	0.0064	**0.0056**	**0.0036**
	DSRG	0.0068	0.0058	0.0050	0.0036	0.0081	0.0070	0.0061	0.0041
	Ours	0.0065	0.0056	0.0049	0.0034	**0.0073**	**0.0063**	0.0057	**0.0036**

Figure 7 and Figure 8 show the ATE and RPE plots that each column contains the results of the same sequence and each row represents the same method. The ATE and RPE plots in the first row are generated by ORB-SLAM2[6]. From the second row to the bottom row show the results of SegNet[12], DSRG[21] and our proposed network[22]. As we can see from the figures, the errors are significantly reduced with the semantic segmentation methods. Even though our semantic segmentation module is weakly supervised, our overall system performance is close and comparable to the SLAM system using fully supervised SegNet.

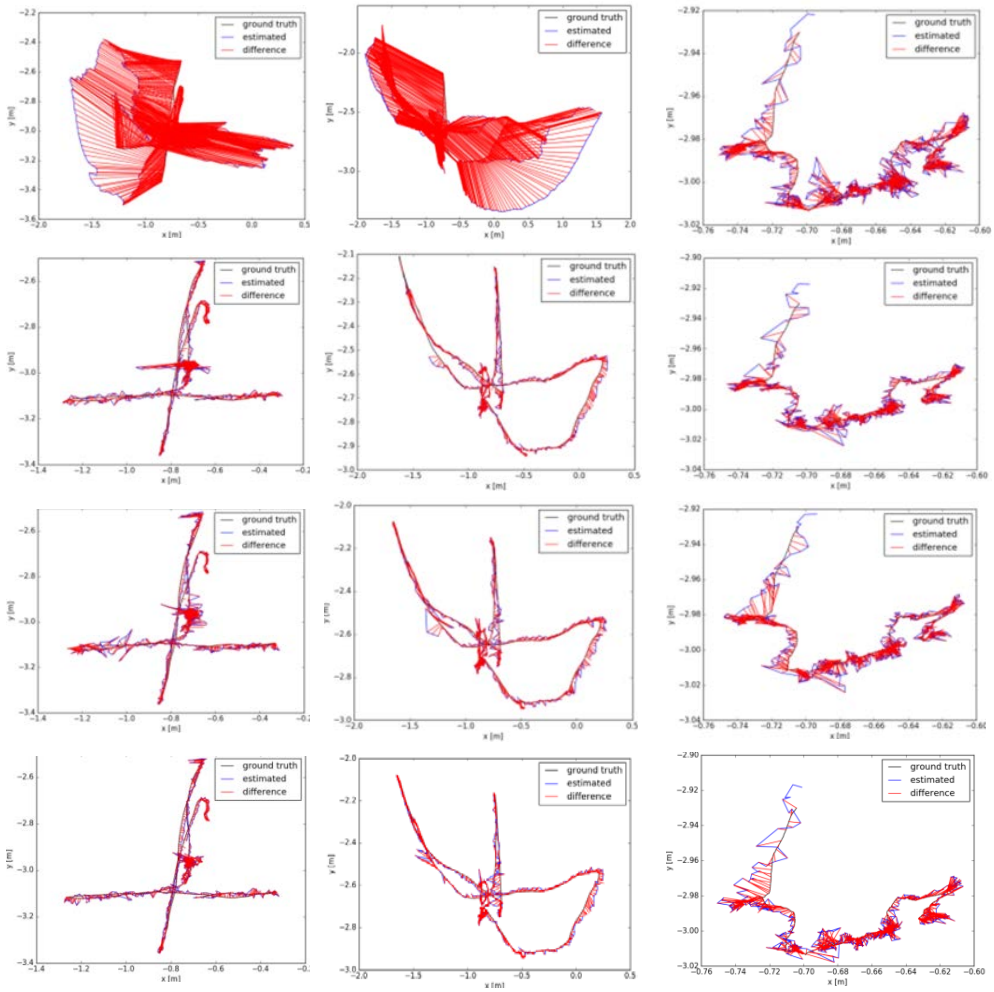
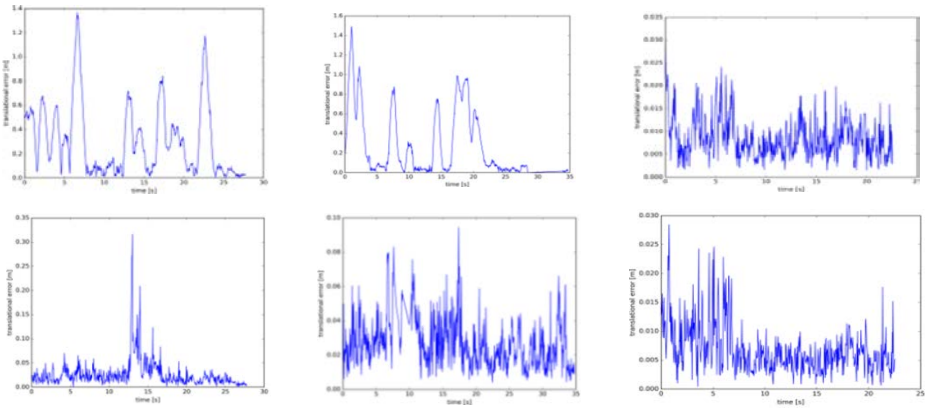


Figure 7. The ATE of sequences (from the left column to the right) xyz, halfsphere and sitting. In each row from top to down shows the ATE generated by : ORB-SLAM2, +SegNet[12], +DSRG[21], +our network[22].



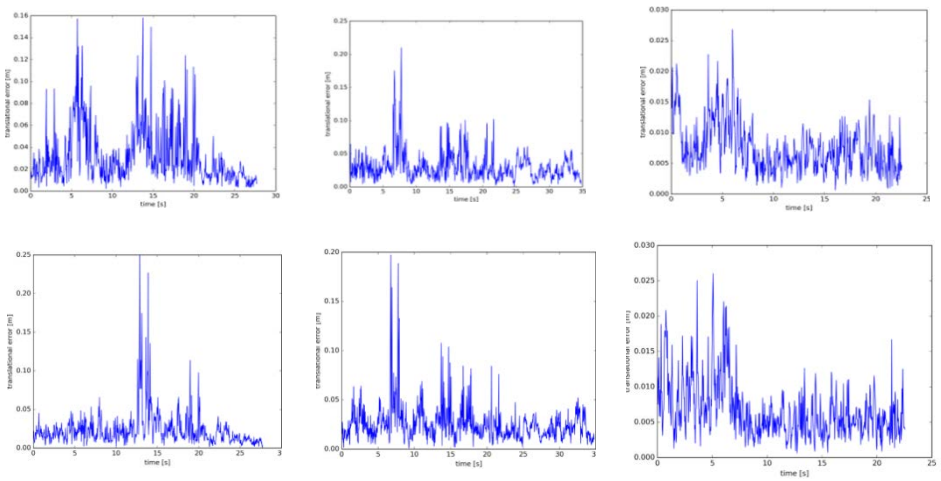


Figure 8. The RPE in the sequences (from the left column to the right) xyz, halfsphere and sitting. In each row from top to down shows the trajectory generated by : ORB-SLAM2, +SegNet[12], +DSRG[21], +our network[22].

5.4 Time cost evaluation

For practical applications, real-time performance is a crucial indicator to evaluate SLAM system. We test the time required for some major modules to process. The results are shown in Table III. The average time for fully supervised model SegNet to process each frame is 36ms, And the average time for the weakly supervised model DSRG and the model in this paper to process each frame in the semantic segmentation thread is 40ms and 41ms, respectively. The weakly supervised semantic segmentation time is slightly longer, because the weakly supervised semantic segmentation captures the contextual information of the image and optimizes the segmentation results using conditional random fields.

TableIV. Time Evaluation

Sequence	**methods**	**Time(s)**		
		mean ORB extract time	**mean movingdetection time**	**mean segmentation time**
xyz	SegNet	0.0094798	0.02063934	0.0358515
	DSRG	0.0100359	0.02064904	0.0406992
	Ours	0.0103220	0.02096500	0.0417910
halfsphere	SegNet	0.0091383	0.0189226	0.0362912
	DSRG	0.0096011	0.0192021	0.0404659
	Ours	0.0101140	0.0162610	0.0414190
sitting	SegNet	0.0091383	0.0189226	0.0362912
	DSRG	0.0096011	0.0192021	0.0404659
	Ours	0.0099390	0.0195530	0.0418610

6. CONCLUSION

In this paper, we propose to use the weakly supervised semantic segmentation method to improve the accuracy and robustness of SLAM system, which solves the problem of using expensive annotation in training for fully supervised semantic segmentation network. The weakly supervised semantic segmentation network is combined with moving consistency check to filter out dynamic targets of the scene. Then the dynamic feature points would be removed, and thus improve the performance of visual SLAM system in dynamic scenes. Experimental results on TUM RGB-D dataset demonstrate that our method improves the accuracy and robustness of the SLAM system.

ACKNOWLEDGEMENT

This paper is granted under the proposal of Chunhui project of the Ministry of Education (No Z2017084) and Plan of Scientific and technological office of Chengdu (No 2015-NY02-00336-NC).

REFERENCES

[1] Cheeseman P, Smith R, Self M. A Stochastic Map for Uncertain Spatial Relationships[C]//4th International Symposium on Robotics Research. 467-474 (1987).

[2] Gao X, Zhang T, Yan Q R, et al. 14 lectures on visual SLAM: From theory to practice[M]. Beijing: Publishing House of Electronics Industry, (2017).

[3] Davison A J, Reid I D, Molton N D, et al. Mono SLAM: real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 29(6): 1052-1067 (2007).

[4] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces[C]//2007 6th IEEE and ACM international symposium on mixed and augmented reality. IEEE, 225-234 (2007).

[5] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-Scale Direct Monocular SLAM[C]//European Conference on Computer Vision, Springer, Cham, 834-849 (2014).

[6] Forster C, Zhang Z, Gassner M, et al. SVO: Semidirect visual odometry for monocular and multicamera systems[J]. IEEE Transactions on Robotics, 33(2): 249-265 (2016).

[7] Mur-artal R, Tardos J D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras [J]. IEEE Transactions on Robotics, 33(5): 1255-62 (2017).

[8] Alcantarilla P F，Yebes J J，Almaza J，et al． On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments [C]. IEEE International Conference on Robotics and Automation, 1290-1297 (2012) ．

[9] Gao C Q, Zhang Y Z, Wang X Z. et al. Semi-direct RGB-D SLAM Algorithm for Dynamic Indoor Environments[J]. Robot, 41(03):372-383 (2019).

[10] Lin Zhilin, Zhang Guoliang, Yao Erliang, etc. Stereo vision odometer based on moving object detection in dynamic scene [J] (in Chinese). Journal of Optics, 37 (11): 187-195 (2017).

[11] Zhang Hexin, Xu Hui, Yao Erliang, Song Haitao, Zhao Xin. A Robust Stereoscopic Mileage Calculation Method in Dynamic Scenes [J] (in Chinese). Journal of Instrumentation, 39 (09): 246-254(2018).

[12] Yu C, Liu Z, Liu X J, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 1168-1174 (2018).

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEEtransactions on pattern analysis and machine intelligence, 2481–2495 (2017).

[14] Bescos B, Facil J M, Civera J, et al. DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes [J]. IEEE Robotics and Automation Letters, 3(4): 4076-83 (2018).

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2961–2969 (2017).

[16] Xi Zhihong, Han Shuangquan, Wang Hongxu. Synchronous localization and semantic mapping of indoor dynamic scenes based on semantic segmentation [J] (in Chinese). Computer Applications, 39 (10): 2847-2851 (2019).

[17] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network [M]. 30th Ieee Conference on Computer Vision and Pattern Recognition.6230-9(2017).

[18] Xiao L, Wang J, Qiu X, et al. "Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment,"Robot. Auton. Syst., vol. 117, pp. 1–16, Jul. (2019).

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C.Berg, "Ssd: Single shot multibox detector," in European conference oncomputer vision. Springer, 21–37 (2016).

[20] Kolesnikov A，Lampert C H．Seed expand and constrain: three principles for weakly-supervised image segmentation[C]//Proceedings of European Conference on Computer Vision. 695–711 (2016).

[21] Huang Z, Wang X, Wang J, et al. Weakly-Supervised semantic segmentation network with deep seeded region growing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7014-7023 (2018).

[22] Yao Q, Gong X. Saliency Guided Self-Attention Network for Weakly and Semi-Supervised Semantic Segmentation[J]. IEEE Access, 8: 14413-14423(2020).

[23] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2921-2929 (2016).

[24] Jiang H, Wang J, Yuan Z, et al. Salient object detection: A discriminative regional feature integration approach[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2083-2090 (2013).

[25] Adams R, Bischof L. Seeded region growing[J]. IEEE Transactions on pattern analysis and machine intelligence, 16(6): 641-647 (1994).

[26] Krähenbühl, Philipp, Koltun V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials[J]. (2012).

[27] Jürgen Sturm, Engelhard, Endres F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 573-580 (2012).

[28] Sun T, Sun Y, Liu M, et al. Movable-object-aware visual slam via weakly supervised semantic segmentation[J]. arXiv preprint arXiv:1906.03629 (2019).

[29] Dai J, He K, Sun J. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 1635−1643 (2015).

[30] Bearman A, Russakovsky O, Ferrari V, Li FF. What's the point: Semantic segmentation with point supervision. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 549−565 (2016).

[31] Lin D, Dai J, Jia J, He K, Sun J. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 3159−3167 (2016).