Combining ElasticFusion with PSPNet for RGB-D based Indoor Semantic Mapping

Weiqi Wang
Zhengzhou Institute of Surveying and Mapping
Zhengzhou, China
809741461@qq.com

Jian Yang
Zhengzhou Institute of Surveying and Mapping
Zhengzhou, China
jyangtum@qq.com

Xiong You
Zhengzhou Institute of Surveying and Mapping
Zhengzhou, China
youarexiong@163.com

***Abstract*—Semantic segmentation has been a research focus of scene parsing and robotic manipulation. And increasing efforts are being made to empower geometry-solely robotic mapping with semantic sensing ability in complex spatial cognition tasks, the so-called semantic mapping. However, semantic mapping has not yet fully exploited the state-of-the-art of semantic segmentation and suffered limitations of application scenario, number of object category and accuracy of object detection. Towards a robust semantic mapping solution, we propose a RGB-D based fusion method that combines an improved ElasticFusion with PSPNet. PSPNet has the potential for breaking restrictions of limited object category in semantic segmentation of RGB images. As for map rendering, the improved ElasticFusion algorithm is validated for its robust performance in indoor scenes in terms of overall model accuracy and integral object shape geometry. Based on the two aforementioned modules, we fuses the semantic images and the depth images to generate 3D point cloud with semantic information. Experiments on the ICL-NUIM datasets have proven the feasibility of the proposed method.**

***Keywords*—semantic mapping, ICP algorithm, ElasticFusion**

I. Introduction

Maps are essential tools for robots to understand and reason about the surrounding environments [1]. However, making maps for robots is still a challenging research pursuit. With the considerable progress in the field of SLAM (Simultaneous Localization And Mapping), robots can map the environment with sparse or dense point cloud and navigate with 2D or 3D occupancy grid maps [2]. The maps made by robots usually contain rich and accurate geometric information but lack semantic information. Without built-in semantics in the map, it would become much difficult or even impossible for a robot to understand its surroundings and perform complex tasks. In order to improve robots' autonomy, it's desirable to equip maps with semantic concepts of corresponding geometric entities, namely semantic mapping. Maps that integrate both semantic and geometric information promote robots' abilities from path planning to task-driven planning. At present, the robotic semantic mapping researches are mostly performed as a objects' classification problem, which is at the bottom of the pyramid of semantic mapping [2].

Semantic mapping can be realized using a framework that consists of two modules, semantic segmentation performed by SVM or CNNs and map rendering performed by SLAM algorithms using RGB-D, monocular or stereo sensors [3,4]. Semantic segmentation is to label the relevant information of objects and currently it mainly focuses on objects' classifications. The advent of FCNs (Fully Convolutional Networks) [5] has significantly boosted the accuracy of semantic segmentation and enabled to handle input images of arbitrary sizes. A variety of convolutional neural networks have been developed for semantic segmentation thereafter, while achieving improvements based on the FCNs' structure [6,7,8,9]. In these advanced networks, the Pyramid Scene Parsing Network (PSPNet) [6] achieves prime performance by incorporating robust global features and proposing an optimization strategy with deeply supervised loss. Therefore, PSPNet is utilized for semantic segmentation for its state-of-the-art performance.

As for map rendering, it integrates semantic information with geometric information to make 2D grid maps or 3D point cloud maps. With rich geometric and texture information, dense 3D point cloud provides a better scene representation than 2D maps and sparse 3D point cloud maps. Dense point clouds are usually generated by 3D laser scans or RGB-D sensors. And mapping solutions based on RGB-D sensors benefit from their economic, portability and efficiency concerns. ElasticFusion [10] is more suitable for representing semantic information compared with other mature SLAM algorithms using RGB-D sensors, such as RGB-D SLAM [11], Kintinuous [12] and BundleFusion [13]. Because ElasticFusion algorithm uses surfels to generate and fuse point clouds, and proposes a deformation graph to ensure a globally consistent map during loop closures. The advantage of using surfels is that surfels are fit for classifying point clouds and parsing semantic information. Therefore, ElasticFusion algorithm is used to perform map rendering.

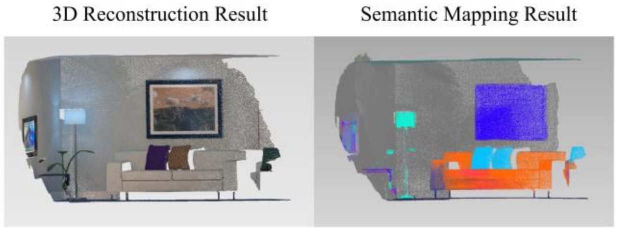


Fig. 1. The output of our method: On the left,mapping result with dense point clouds on the ICL-NUIM lr_kt0 dataset by improved ElasticFusion only. On the right, semantic mapping on the same scene.

Recent developments of semantic mapping focus on two aspects, one is to improve the accuracy of semantic segmentation to improve the quality of semantic maps, and the other is to utilize semantic information in the looping closure detection module in SLAM to get more consistent maps. The improvements can be found in the literatures [4,10,14]. [14] discussed the application of semantic maps but only considered few object classes in the semantic segmentation. [4] extracted semantic information of simple objects and ORB features to assist localization, in which the map representation is based on sparse point clouds map. [10] applied probabilistic multiplication to perform semantic mapping by mapping semantic segmentation results onto dense 3D point clouds. The semantic segmentation employed a probabilistic event model, which failed to generate accurate semantic maps by probability multiplication. As such, this work aims to improve the accuracy of semantic maps by directly generate semantic maps from the outcome of semantic segmentation instead of mapping the results of semantic results onto 3D point clouds by probability multiplication. In our semantic mapping proposal, PSPNet acts as a front-end for semantic segmentation of RGB images and ElasticFusion algorithm is adopted as a back-end to generate semantic maps through semantic images and depth images. The ICL-NUIM dataset is chosen to verify the feasibility of our method, as shown in Figure 1. The semantic maps can well represent the categories of objects while ensuring the accuracy of their geometric properties.

In the remainder sections, we first discuss related work and then introduce our method in detail. Experiments on ICL-NUIM datasets is discussed in Section IV and the final section for conclusion.

II. Related Work

The related work of semantic segmentation and semantic map rendering are introduced respectively. Semantic segmentation has been improved considerably since FCNs [5] was proposed. FCNs has three major contributions that leads to performance improvement, including using a 1×1 convolutional layer to represent spatial information, adopting a transposed layer to retain the original resolution of input images and proposing the conception of skip connection to boost semantic segmentation robustness. Based on FCNs, many improvements are achieved through adding more layers to the networks or fine-tuning the structure of the networks, such as Deeplab [7], ICNet [8] and SegNet [9]. Besides the structural modification, some researchers made full use of the CRF and MRF to improve the performance of semantic segmentation [15,16,17]. And [18] adopted Recurrent Neural Network and CRF to perform semantic segmentation. Comparing with the above neural networks, PSPNet [6] obtains better results in either indoors or outdoors environments.

Rendering semantic maps for robots is a challenging task in semantic SLAM. With the development of deep learning, semantic labelling mostly relies on CNNs or RNNs and departs from SVM. Researchers attempts to achieve object detection and semantic labelling using different sensors in SLAM system [3,19,20,21]. [4] presented the concept of probabilistic data association to help localization and looping closure detection in SLAM by using semantic information. This method significantly improved the accuracy of estimated trajectories but only produce maps with simple representation of geometry and texture. [22] focued on labelling more categories of objects and improving the accuracy of semantic segmentation on each frame image. [19] emphasized on the use of semantic maps which only represented a few objects. [10] adopted a mapping method of mapping semantic annotation results into 3D reconstruction point clouds to gain 3D semantic annotated maps by probability multiplication, but the edges of the objects were fuzzy.

There are two better performing frameworks in semantic mapping. One is performing semantic segmentation in front-end and mapping the semantic segmentation results onto 3D point clouds in back-end, such as SemanticFusion. The other is mapping with semantic information and visual features directly, such as [4]. Our method is most related to SemanticFusion [10], using mature convolutional neural networks and ElasticFusion algorithm. We combined the semantic segmentation results with depth images as inputs to generate 3D semantic point clouds maps instead of mapping the semantic segmentation results to 3D point clouds maps by probability multiplication in [10].

III. Method

There are two modules in our semantic mapping pipeline (see Figure 2), a Convolutional Neural Network for semantic segmentation, PSPNet, and a dense visual SLAM system, ElasticFusion. As for semantic segmentation in either indoor or outdoor environments, PSPNet provides accurate semantic segmentation results for each RGB image, which consists of three channels, R, G, and B. Because of the combination of many RGB images and depth images, point cloud can be generated. Then for map rendering, ElasticFusion takes results of semantic segmentation and original depth images as inputs to make semantic maps. In the following two sections, two modules are discussed in detail.

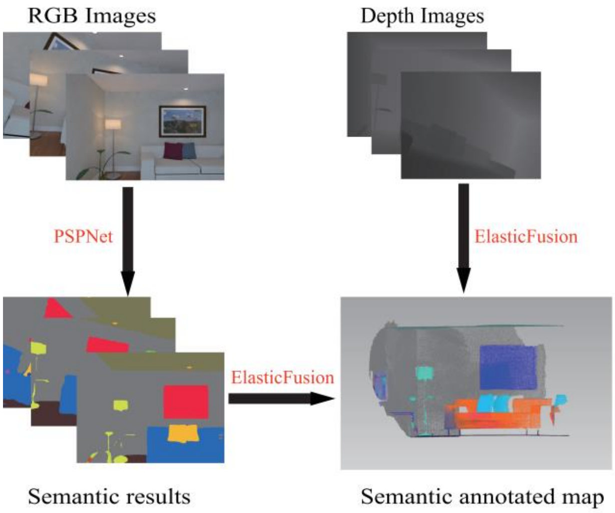


Fig. 2. An overview of our method

A. Semantic Segmentation using PSPNet

PSPNet introduced a pyramid pooling module, which served as a expressive contextual prior that was capable of extracting global contextual information for semantic segmentation and scene parsing. For example, it's less likely to label a boat in the river as a car when taken the context river into account [6]. While retaining the advantages of FCNs, PSPNet improves the accuracy of semantic segmentation through incorporating multilevel information provided by a pyramid pooling module. The network architecture of PSPNet is shown in Figure 3 [6]. The original RGB images are taken as CNNs' input to get feature maps, which are important to the pyramid pooling module. After the processing of the pyramid pooling module, concatenating pooling results to do decoding and up sampling, the segmentation results are obtained.

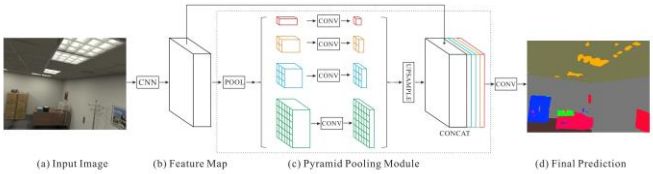


Fig. 3. The PSPNet architecture [6]

Different from SemanticFusion[10] which uses four channel images (RGB and Depth), our semantic segmentation process only requires RGB images and does not rescale the input images. Due to the lack of labelled depth images for model training, SemanticFusion converts the 0-255 color range to 0-8m depth range by increasing the weights [10]. This operation might sabotage the precision of semantic segmentation since the underlying assumptions of perspective distortion does not always hold. For instance, there are the cases of dislocation that make it difficult to obtain accurate depth. Therefore, in our method, depth images are used to generate point clouds only and RGB images are used to make semantic segmentation in order to ensure the accuracy of semantic segmentation.

B. *Map rendering using improved ElasticFusion*

The ElasticFusion algorithm consists of four steps: transforming RGB images and depth images into point clouds and obtaining the coordinates and normal vectors of point clouds, estimating pose parameters by ICP algorithm and photometric method to perform point clouds registration, using random ferns algorithm to achieve loop closure detection, integration and updating of point clouds. ElasticFusion algorithm uses surfel-based representation to model observed scenes, it is more suitable for semantic labelling. However, there are some shortcomings in the process of using ElasticFusion algorithm, such as poor model quality of local point clouds in certain objects (see Figure 4). Therefore, we make minor improvements to the algorithm by altering the strategy of searching matching points in ICP algorithm.

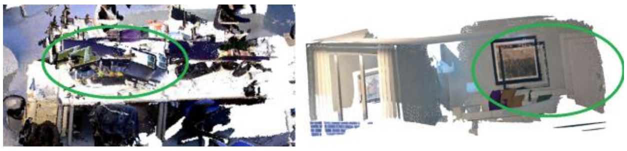


Fig. 4. Illustration of poor model quality of local point clouds

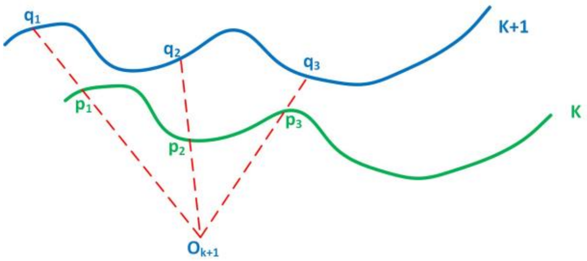


Fig. 5. Illustration of search of the matching points of the given pose point using projection method

Original ICP algorithm search the matching points by using projection method, as shown in Figure 5. And ICP algorithm solves the optimal pose parameters iteratively, while the matching relationships in the point pairs remain constant during the iteration process. Therefore, whether the matching points are selected properly or not, it directly undermines the accuracy of the pose parameters and thus the reconstruction results of the point cloud model. Our method introduces a new point searching strategy to the projection-baesd matching points selection. A circular search region Q with a certain radius is first drawn, and then all candidate points containing the frame point cloud are traversed to find the best matching point, as shown in Figure 6.

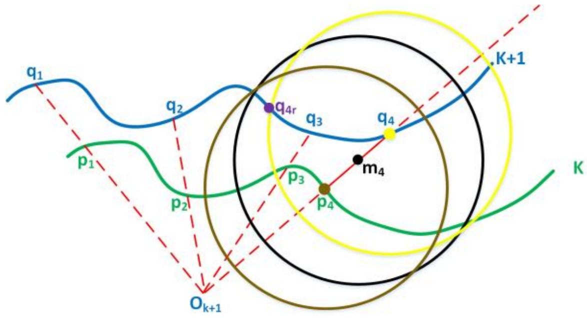


Fig. 6. Illustration of search of the matching points using cicles with different centers

For instance, to search a matching point of p_4, all points contained in the circular search area are candidates $q_i(q_i \in Q)$. The optimal search can be formulated as the solution of the equation

$$f(q_i) = \lambda \cdot N(q_i) \cdot N(p_4) \tag{1}$$

where λ is an adjustable weight, $N(*)$ is the normal vector of point $*$. The selected match point q^* can be denoted as:

$$q^* \leftarrow max\big(\lambda \cdot N(q_i) \cdot N(p_n)\big) \tag{2}$$

According to the situation of including candidate points, the black center and circle are selected. The center is the midpoint of the line segment and the radius of the circle is 1.5 times the length of the line segment, as shown in Figure 6. When an appropriate initial matching point cannot be selected according to the projection method, either the starting point q_0 or the ending point q in the frame point cloud is selected to determine the radius $r = \frac{3}{4}|p_n q|$ or $r = \frac{3}{4}|p_n q_0|$ of the circular search area, as shown in Figure 7.

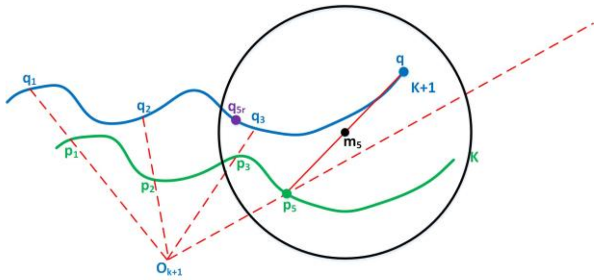


Fig. 7. Illustration of search of the matching points without an initial matching point

IV. Experiments

To testify the feasibility of our method, we evaluate it in this section in two parts, including the quality of reconstruction based on improved ElasticFusion algorithm and the results of semantic mapping. The experiments were implemented on a laptop with an Intel Core i7-7820HK 2.90 GHz CPU and an Nvidia GTX 1070 GPU. The datasets we used are ICL-NUIM datasets, which provided by the Imperial College London and the National University of Ireland Maynooth jointly. The datasets are mainly used to evaluate visual odometry, indoor 3D reconstruction and SLAM system.

A. *Improved ElasticFusion Algorithm*

We used dyson_lab and ICL-NUIM datasets to verify the performances of our method on model reconstruction. In this part, we evaluate the quality of the reconstruction (Figure 8 and 9) and camera trajectory estimation respectively (Figure 10), and illustrate two experimental results briefly (the left image shows the result of the original algorithm and the right image shows the result of the improvement). Figure 8 shows the comparison results on dyson_lab data, Figure 9 and 10 show the comparison of reconstruction results and camera trajectories (red is the ground truth of the camera trajectories, green is the original algorithm's camera trajectories estimation, blue is the improved algorithm's camera trajectories estimation in Figure 10).

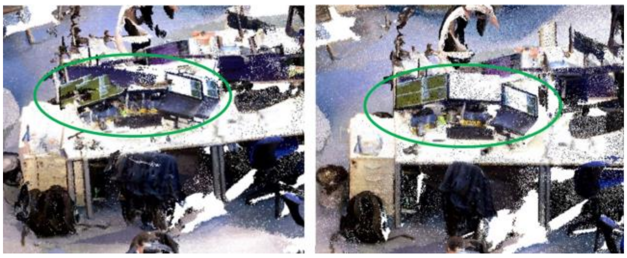


Fig. 8. The comparison results on dyaon_lab data

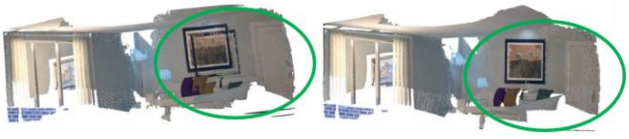


Fig. 9. The comparison results on lr_kt3 data

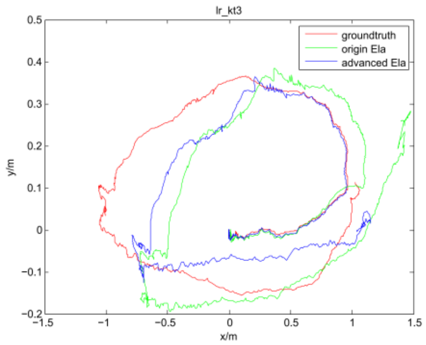


Fig. 10. The comparison charts of the camera trajectories on lr_kt3 data

The improved algorithm reduces the time of final loop closure detection and the probability of loop failure while the original algorithm has a certain probability that the failure of loop closure detection results in poor reconstruction effect in Figure 8. The experimental results demonstrate that the accuracy of pose parameter estimation and the quality of model reconstruction could be significantly improved by modifying the strategy of the matching point selection in ICP algorithm, which serve as a solid basis for accomplishing semantic mapping.

B. *Semantic mapping*

We utilized PSPNet in TensorFlow to implement semantic segmentation on RGB images. We first trained the networks for over 1 day, and performed semantic segmentation results for over 5 hours on using each ICL-NUIM dataset. Then we fused the results with depth images to generate semantic maps using improved ElasticFusion algorithm. As shown in Figure 11, SemanticFusion [10] failed to produce integral object labels. Yet our method manages to generate complete shape geometries of the labeled objects. Figure 1 and 12 show the experimental results on two scenes in the datasets, one for living room datasets (Figure 1， 12 (a) and (b)) and the other for office room datasets (Figure 12 (c) and (d)). The left images illustrate the reconstruction results by the improved ElasticFusion, the middle images are the results of semantic mapping by SemanticFusion and the right ones show the results of semantic mapping in our method.

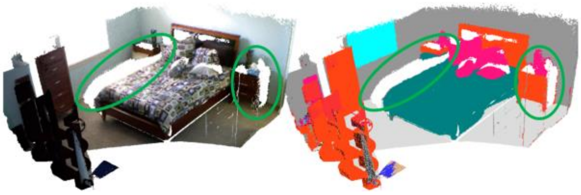
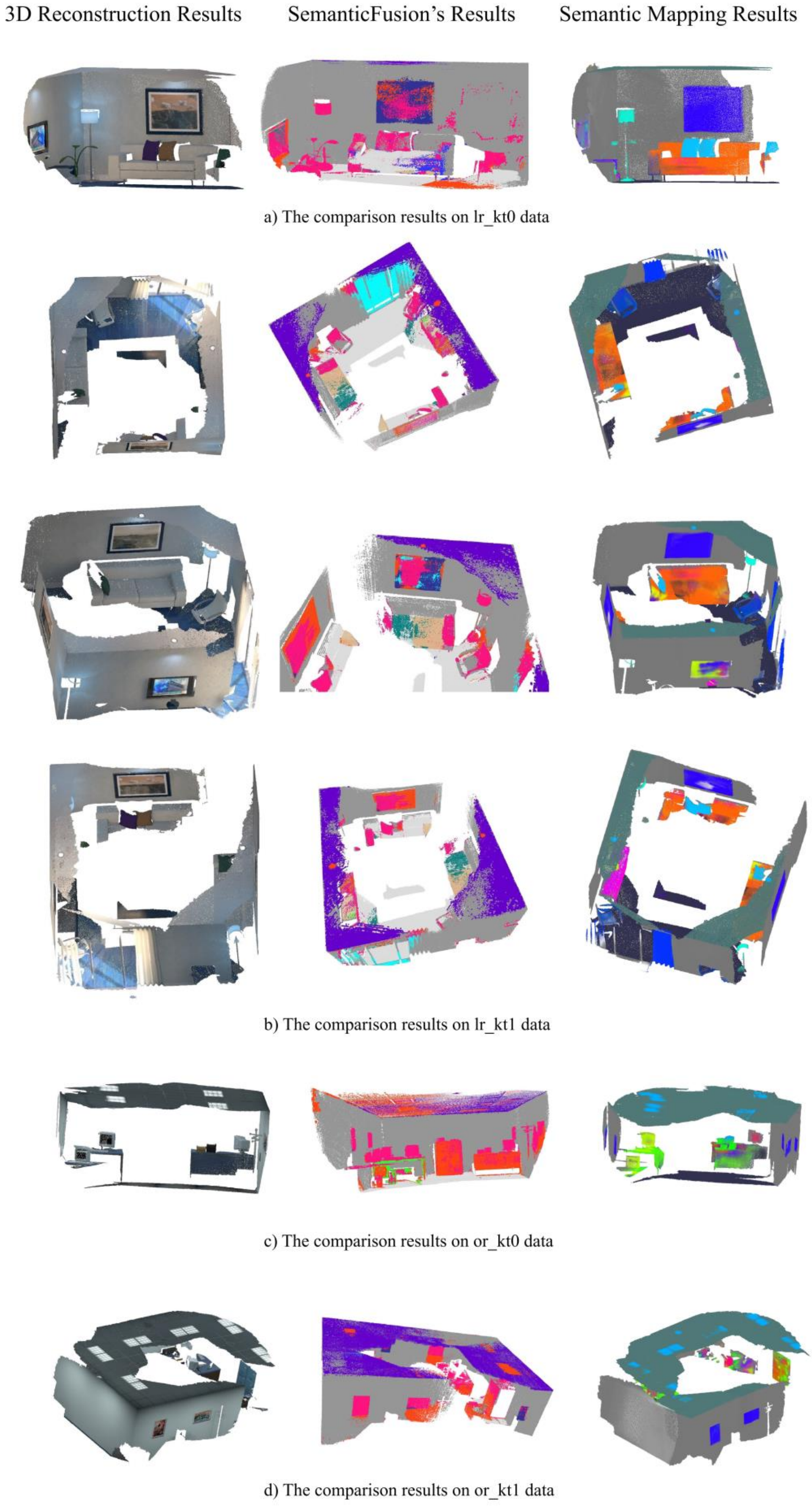


Fig. 11. SemanticFusion's result [10]



a) The comparison results on lr_kt0 data

b) The comparison results on lr_kt1 data

c) The comparison results on or_kt0 data

d) The comparison results on or_kt1 data

Fig. 12. The comparison results on serveal ICL-NUIM data

Experimental results demonstrate that with high accuracy outcome of semantic segmentation on each image, our method could obtain a high quality semantics map and each object has clearly edges and better details. But our method is not as good as SemanticFusion to reconstruct the geometric structure of the room, as shown in the figure 12 (c) and (d). The reason is the deviation of semantic segmentation results from the same object on multi images, as shown in figure 13. The next step is to improve the accuracy of semantic segmentation by introducing high order CRF.

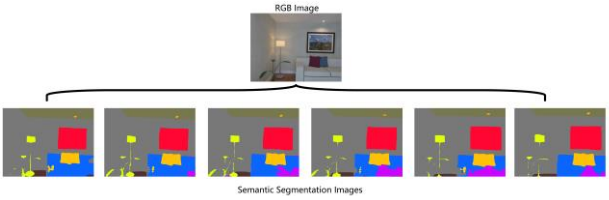


Fig. 13. The deviation of semantic segmentation results

V. Conclusions

We fused semantic segmentation images with original depth images, which takes full advantage of PSPNet and improved ElasticFusion algorithm respectively, to obtain semantic maps. The early results on dyson_lab and ICL-NUIM datasets have shown that the proposed method could obtain better 3D reconstruction models and accurate semantic labels. Since the method is too computational expensive, it can only work in an off-line mode. Thus, it’s desirable to develop an on-line version for real-time semantic mapping.

Acknowledgment

This work is supported by the National Key R&D Program of China (No.2017YFB0503500) and the Project of Science and Technology Innovation of Henan Province (No.142101510005).

References

[1] Thrun Sebastian. Robotic mapping: A survey. Exploring artificial intelligence in the new millennium. 2002 Feb;1(1-35):1.

[2] Cadena C, Carlone L, Carrillo H, et al. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age[J]. IEEE Transactions on Robotics, 2016, 32(6):1309-1332.

[3] Li X, Belaroussi R. Semi-Dense 3D Semantic Mapping from Monocular SLAM[J]. 2016.

[4] Bowman S L, Atanasov N, Daniilidis K, et al. Probabilistic data association for semantic SLAM[C]// IEEE International Conference on Robotics and Automation. IEEE, 2017:1722-1729.

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4):640-651.

[6] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[J]. CVPR, 2016:6230-6239.

[7] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(4):834-848.

[8] Zhao H, Qi X, Shen X, et al. ICNet for Real-Time Semantic Segmentation on High-Resolution Images[J]. 2017.

[9] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2481-2495.

[10] Mccormac J, Handa A, Davison A, et al. SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks[J]. 2016:4628-4635.

[11] Henry P, Krainin M, Herbst E, et al. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments[J]. International Journal of Robotics Research, 2014, 31(5):647-663.

[12] Whelan T, Kaess M, Fallon M, et al. Kintinuous: Spatially Extended KinectFusion[J]. Robotics & Autonomous Systems, 2012, 69(C):3-14.

[13] Dai A, Izadi S, Theobalt C. BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface re-integration[J]. Acm Transactions on Graphics, 2017, 36(4):76a.

[14] Sünderhauf N, Pham T T, Latif Y, et al. Meaningful maps with object-oriented semantic mapping[C]// Ieee/rsj International Conference on Intelligent Robots and Systems. IEEE, 2017:5079-5085.

[15] Baque P, Bagautdinov T, Fleuret F, et al. Principled Parallel Mean-Field Inference for Discrete Random Fields[C]// Computer Vision and Pattern Recognition. IEEE, 2016:5848-5857.

[16] Krähenbühl P, Koltun V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials[J]. 2012:109-117.

[17] Teichmann M T T, Cipolla R. Convolutional CRFs for Semantic Segmentation[J]. 2018.

[18] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional Random Fields as Recurrent Neural Networks[C]// IEEE International Conference on Computer Vision. IEEE, 2016:1529-1537.

[19] Ma L, Stückler J, Kerl C, et al. Multi-View Deep Learning for Consistent Semantic Mapping with RGB-D Cameras[J]. 2017.

[20] Hermans A, Floros G, Leibe B. Dense 3D semantic mapping of indoor scenes from RGB-D images[C]// IEEE International Conference on Robotics and Automation. IEEE, 2014:2631-2638.

[21] Hosseinzadeh M, Latif Y, Pham T, et al. Towards Semantic SLAM: Points, Planes and Objects[J]. 2018.

[22] Morris A, Danial W, Johann P, et al. Multi-view 3D Entangled Forest For Semantic Segmentation and Mapping[C]// IEEE International Conference on Robotics and Automation (ICRA). 2018