Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

A Dense Segmentation Network for Fine Semantic Mapping

Guoyu Zuo, Tao Zheng, Zichen Xu and Daoxiong Gong
Faculty of Information Technology
Beijing University of Technology
Beijing 100124, China
zuoguoyu@bjut.edu.cn

Abstract— This paper proposes a fine semantic mapping method using dense segmentation network (DS-Net) to obtain good performance of semantic mapping fusion, in which the semantic segmentation network (DS-Net) is constructed based on the idea of DenseNet's dense connection. First, the RGB image and the depth image are used to generate a dense indoor scene map via the state-of-the-art dense SLAM (ElasticFusion). Then, semantic segmentation are precisely performed on the input RGB image via DS-Net. Finally, the long-term correspondence between the landmarks and the indoor scene map is established using the continuous frames both in the visual odometer and loop detection, and the final fused semantic map is obtained by integrating semantic predictions of the RGB-D video frames of multiple angles with the indoor scene map. Experiments were performed on the NYUv2 and CIFAR10 datasets and our laboratory environments. Results show shows that our method can reduce the error of dense map construction and obtain good semantic segmentation performance.

Index Terms— Semantic segmentation, RGB-D, DS-Net, DenseNet, NYUv2

I. Introduction

In the fields of robotics and computer vision, semantic map lays a foundation for realizing human-robot interaction and human-robot fusion, and it can be widely applied in robot navigation, robot manipulation and augmented reality. How to construct an incremental and robust semantic map in real time has always been an important research issue. Due to the rapid development of simultaneous localization and mapping (SLAM), robot can use sparse or dense point clouds to map environments and navigate in 2D or 3D grid maps [1], and great achievements have been made in autonomous navigation and automatic obstacle avoidance by using geometric map and feature map. But the robot agent cannot get more information from the maps which contain geometric and point cloud information. If these maps don't possess semantic information, it is difficult and even impossible for robot to understand the complex environment information. Semantic map that integrates semantic and geometric information is needed to be established between object and semantics, and finally to improve the ability of robot agent in path planning and other more sophisticated tasks.

Currently, semantic mapping is mainly based on a framework composed of two modules: semantic segmentation performed by convolutional neural network (CNN) and map construction based on SLAM. Some CNN methods focused on improving the accuracy of semantic segmentation [2], [3]. However, to maximally extract all the information in the map, we often deepen the network layers and build the robot system more computationally rigorously, because the operations, such as 3D reconstruction, camera pose estimation and CNN-based semantics require a large number of computing resources. To achieve real-time performance, McCormac et al. [2] constructed a semantic 3D map by combining CNN and the dense SLAM system, and Hermans et al. [4] proposed a novel 2D-3D label transfer based on Bayesian updates and dense pairwise 3D Conditional Random Fields. Although the above frame skipping strategy can improve their run-time performance, it limits their application, since it tends to bring in inaccuracy in fast camera motions.

The SLAM system provides the correspondence from 2D frames to globally consistent 3D maps. Compared with other mature SLAM systems such as RGB-D mapping [5], Kintinuous [6] and BundleFusion [7], ElasticFusion [8] is more suitable for representing semantic information. ElasticFusion algorithm uses surfels to generate and fuse point clouds, and deformation maps are used to ensure globally consistent mapping during closed loop. The advantage of surfels is that they are suitable for classifying point clouds and parsing semantic information.

To construct real-time and efficient semantic map, we propose a semantic segmentation network based on DenseNet [9] by combining the dense SLAM system ElasticFusion and dense segmentation network (DS-Net), in which 2D segmentation of input frame is performed by

*This work is partially supported by National Natural Science Foundation of China (61873008) and Beijing Natural Science Foundation (4182008).

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

using the dense segmentation network (DS-Net) model we designed, and the mapping from 2D segmentation to a 3D point clouds is realized by using the Bayesian framework. In our semantic mapping proposal, DS-Net has more simplified network structure and more efficient real-time performance than the RGB-CNN proposed by Noh et al. [10] when ensuring the accuracy of semantic segmentation. We selected the CIFAR10, NYUv2 dataset [11] and the real environment of the laboratory to verify the feasibility of our approach.

II. Background

A. Semantic mapping

Semantic mapping is a challenging task for robots in semantic SLAM, its main work is to construct a dense, semantically annotated 3D map for indoor scenes. Some work identified only partial 3D maps without generating dense semantic 3D maps [12], [13]. Bowman S L et al. [12] improved the performance of RGB SLAM in camera pose and scale estimation by utilizing not only low-level geometric features such as points, lines, and surfaces but also using the detected targets as landmarks. Salas-Moreno et al. [13] mapped indoor scenes at the level of semantically defined objects, but this method is limited to mapping objects in pre-defined databases. It does not provide the dense labeling of the entire scene, which includes walls, floors, doors, and windows. Nakajima Y et al. [14] proposed an efficient semantic mapping approach by assigning class probability to each region of the 3D map which is built through a SLAM framework, a ResNet-based network structure and geometric-based segmentation.

B. 2D semantic segmentation

CNN can greatly reduce the input resolution through successive pooling layers and it is well suited for image classification task and semantic segmentation. The semantic segmentation structure generally consists of a CNN network and deconvolution modules. CNN realizes feature extraction of input image and assigns an initial category label to each pixel. Deconvolution module enable the network to output probability density map whose resolution is the same as that of input image. Long et al. [15] introduced fully convolutional networks (FCN) in semantic segmentation, which greatly improves the segmentation performance. Ivan Krešo et al. [16] put forward a ladder-style segmentation network which can improve the semantic segmentation accuracy to a certain extent, but the network structure of jump connection greatly increases the number of layers and parameters of the network, so it is not suitable for real-time application.

III. Method

As shown in Fig. 1, our method consists of three main parts: The SLAM framework for real-time reconstruction of indoor scenes, a specially designed 2D semantic segmentation network DS-Net, and a Bayesian update scheme. First, a geometric edge map is generated from the current depth frame, and the RGB image and the depth image obtained from the input are used to generate a dense indoor scene map via the ElasticFusion system. Second, the input RGB image performs precise semantic segmentation via DS-Net and returns the class probability of each set of pixels. Finally, the update class probability is assigned to each surfel in the 3D map by the Bayesian update scheme, and these probabilities are updated by using the correspondence between frames provided by SLAM and generate the final semantic map. The details of this method are described as follows.

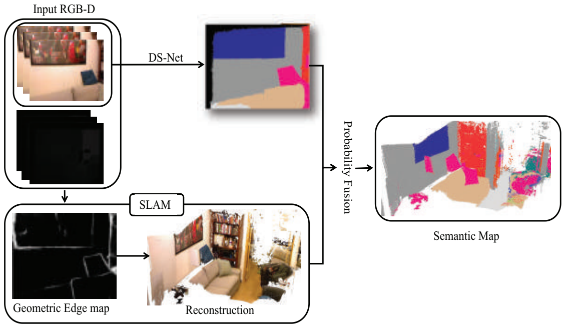

Fig. 1. An overview of our method.

A. SLAM Mapping

In our system framework, ElasticFusion is adopted to produce a dense 3D map. ElasticFusion integrates point cloud information through the surfel model, and updates, matches, displays and projects the point cloud based on OpenGL. The reconstruction accuracy can be achieved by continuously optimizing the reconstructed map. Therefore, modeling the observed scenes using a map based on surfels is more suitable for semantic annotation. The ElasticFusion algorithm consists of four steps: converting RGB images and depth images into point clouds, acquiring coordinates and normal vectors of point clouds, estimating the camera pose parameters by the ICP algorithm and photometric method for point cloud registration, and using the random ferns algorithm to achieve loop detection, integration and point cloud updates.

B. DS-Net Architecture

Fig. 2 shows the DS-Net network structure we designed on the basis of DenseNet. DS-Net consists of four dense

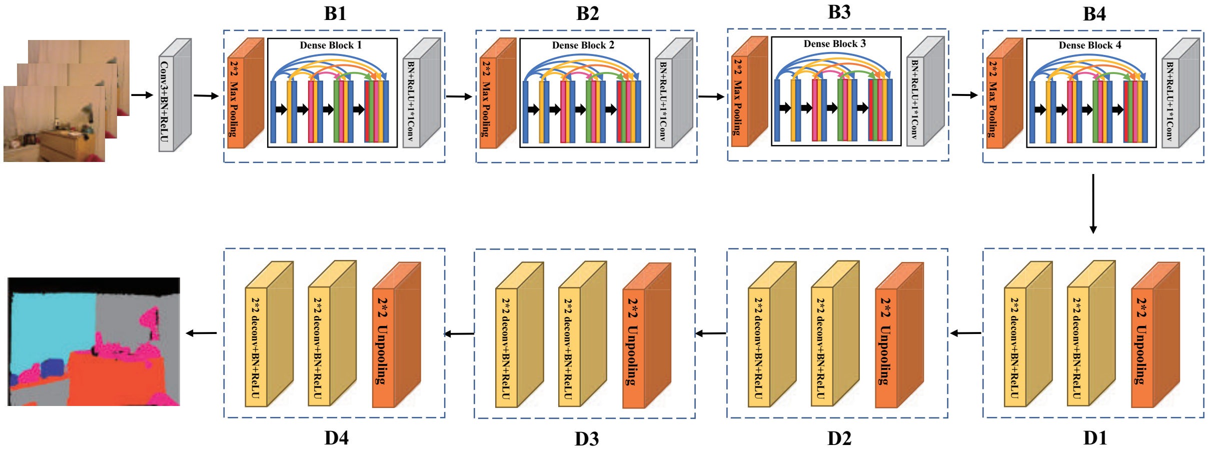

Fig. 2. The network architecture of DS-Net.

blocks, eight deconvolutions, four unpoolings, and other operations, such as batch normalization (BN) and rectified linear units (ReLU). The dashed box B1, B2, B3, B4 consist of a pooling operator, a dense block, and a convolution layer; the dashed box D1, D2, D3, D4 enclose an unpooling operator and two deconvolution layers. The first half of the pooling operator, the dense block, and the convolution layer are used to fully extracts the high-level features and low-level features of the input image by dense connections between dense blocks. The second half of the deconvolution network and the unpooling network are used to recover high-quality images from previously extracted feature maps. Our method generates object segmentation masks using deep deconvolution network, where a dense pixel-wise class probability map is obtained by successive operations of unpooling, deconvolution, BN and ReLU.

1) Dense Block: In Fig. 3, the image is first transmitted into dense block, and the direct connection of different convolution layers is used to fully extract the advanced features and low-level features in the image, and the resolution of the image has not changed. In a dense block, any two adjacent layers are directly connected by multiple operations such as batch normalization (BN), rectified linear unit (ReLU), and convolution layers. A introduction 1×1 convolution layer in B1 is mainly to decrease the number of input feature maps due to too many input features after concatenations, resulting improved computational efficiency. In multiple B modules, the linear growth in the number of features is compensated by the reduction in the spatial resolution of each feature map after the pooling operation. Therefore, each layer in the dense block can not only benefit from the low-level functions and advanced functions in the feedforward setting but also reduce the risk of gradient explosion or disappearance.

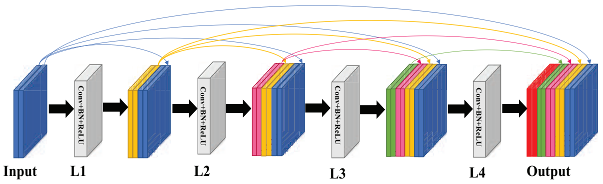

Fig. 3. A dense block with 4 convolution layers.

2) Deconvolution Network: The deconvolution network reconstructs image from the extracted features obtained from the dense blocks. The deconvolution network shown as the dashed boxes D1, D2, D3 and D4 in Fig. 2, which are all composed of the deconvolution and unpooling modules. The deconvolution network upsamples the previous feature map, expands the image pixels, performs deconvolution, and obtains the weights through learning. Since the upsampling path increases the spatial resolution of the feature map, the linear increase in the number of features has a high demand for memory requirement, especially for full resolution features in the pre-softmax layer. To overcome this limitation, we use the crossover structure of unpooling and deconvolution in the deconvolution network. We do not use the skip structure here because it is not a fundamental solution, the performance gain is not significant and will bring a lot of parameters. Therefore, in our network structure, the deconvolution is only connected to the last dense block and not to all previous dense blocks. The feature map the last dense block outputs already contains the information from all previous dense blocks with different resolutions.

C. Probability Fusion

Camera pose based on SLAM can be used to establish the relationship between the pixels of each keyframe and realize incremental fusion by stacking the semantic label information of keyframes with the spatial coordinates of the SLAM landmarks as in SemanticFusion. However, a single 2D semantic segmentation will result in inconsistent labels between successive frames during camera movement. For the keyframe K_t of the input camera at this time, we define the category label distributed on the surfels S_n in the 3D map as l_i. If the spatial coordinates of Sn is not fused in the keyframe K_t, the corresponding category label is $p(l_i|K_{1,\ldots,t})$, and we can use recursive Bayesand standardized constant Z to update this probability:

$$p(l_i|K_{1,\ldots,t}) = \frac{1}{Z_i} p(l_i|K_t) p(l_i|K_{1,\ldots,t-1}) \tag{1}$$

In order to obtain the probability of the corresponding category label on the S_n in the 3D map given the corresponding keyframe, we define the pixel coordinates of the surfels S_n in the keyframe K_t as $x(S_n, t)$, space coordinate matrix of the corresponding category label as $M_{x(S_n,t)}$, and the probability value after recursive Bayes update and standardized constant Z is:

$$p(S_n \to l_i|K_{1,\ldots,t}) = \frac{1}{Z_i} p(M_{x(S_n,t)} = l_i|K_t) p(l_i|K_{1,\ldots,t-1}) \tag{2}$$

IV. Experiments

To evaluate our approach, we use the VOC2012 split dataset to train the proposed deep network, and fine-tune the network on the training set of the NYUv2 dataset for the 13 semantic classes defined by Couprie et al. under the caffe framework. We test the performance of the parameters of the network on the CIFAR10 dataset, and compare our constructed system with SemanticFusion on the NYUv2 dataset and our laboratory environment. The experiment was performed on an Intel Core i7 3.3GHz CPU and Nvidia GTX 1060 GPU.

A. Network Parameter Tuning

The parameters to be optimized in the DS-Net network include the number of convolution layers in each dense block, and the number of feature maps in each dense block. The optimal network parameters are selected according to the experimental effects using different parameters. In experiment, the number of network layers L in the dense block is set to 3, 4 and 5, respectively, and the number of feature maps k in dense block is set to 8, 16 and 32, respectively. The CIFAR10 dataset has a total of 60000 color images, we use 50000 images of them for training, and another 10000 images for testing. We perform an accuracy calculation every 500 iterations for each number of experimental parameter. The average accuracy of 100 test results is used as the experimental result. Fig. 4 and 5 show the results of two parameters, respectively.

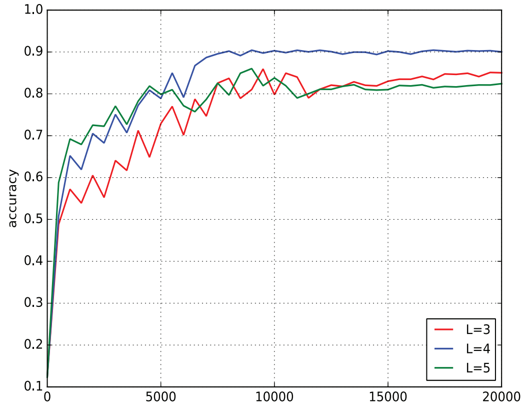

Fig. 4. Accuracy of different number of layers in dense block.

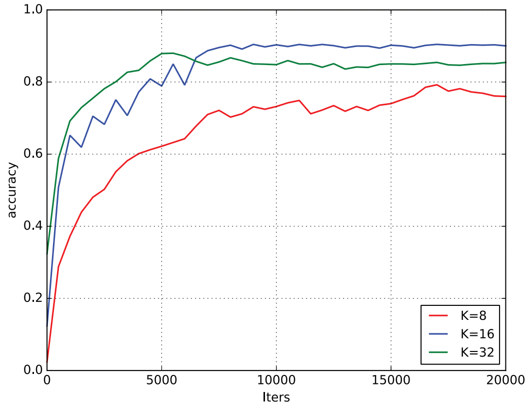

Fig. 5. Accuracy of different number of feature maps in dense block.

From Fig. 4, we can see that when L changes from 3 to 4, the prediction accuracy of the network is significantly improved, and the convergence speed and stability of loss are improved. When L changes from 4 to 5, the accuracy of network prediction of L=5 is improved compared with L=4 in the early time of iterations. However, as the number of iterations increases, there is no more improvement in the prediction accuracy, even lower than the accuracy of L=3. When L=5, the convergence speed of the network's loss becomes slower and unstable than the other two curves. This is because more training samples are needed to avoid overfitting as the depth of the network increases. And the deep network has a serious impact on the memory usage and the real-time calculation because it has more parameters and longer time for network training. Therefore, we choose L=4 for the dense block in our system.

From Fig. 5, we find that the network can achieve the best results when k=16. In the later stage of network training, the reason of lower network accuracy is that the number of feature maps is not enough to achieve full extraction of image features. When k=32, as the network width increases too much, the number of parameters of the network increases significantly, and the convergence becomes more difficult. Therefore, considering both the accuracy and the training time, we choose $k = 16$ as the optimal number of feature maps.

B. Semantic mapping

In this section, we implemented precise semantic segmentation based on RGB image under caffe and compared our improved fusion algorithm with SemanticFusion on NYUv2. The results are shown in Fig. 6, in which the first column are the original RGB images, the second are the dense maps of the indoor environment constructed by ElasticFusion, the third are the semantic segmentation maps obtained by SemanticFusion, and the last are the semantic segmentation maps obtained by our method.

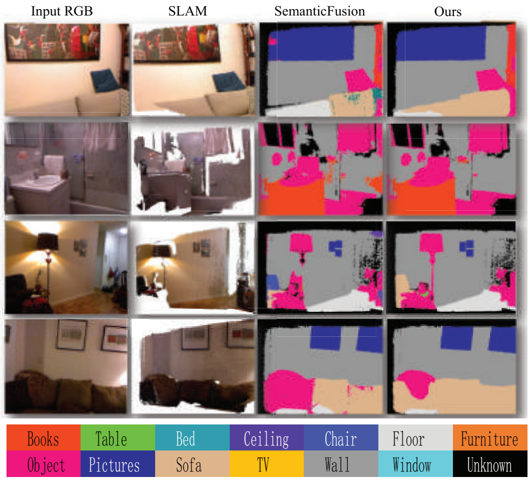

Fig. 6. The results of our algorithm compared with SemanticFusion on the NYUv2 dataset.

For the objects that does not appear completely in the video frame, such as the rightmost pillow in the first scene of Fig. 6, although the ElasticFusion algorithm can realize the reconstruction of the pillow, it has an erroneous probability prediction in the experimental results of SemanticFusion for the pillow. This is because after the Bayesian update, the erroneous probability prediction results are used in the dense map, and the semantic map of the error segmentation is obtained. In our system, because the densely connected network in DS-Net fully extracts the edge features of the object, it can generate more accurate probability prediction maps and get better results. For the third scene, the strong light, which has a great influence on object detection, our segmentation also has a good result., strong light may make the object unable to detect the object from a certain angle of view, but with multiple angles of video frame sequence, we can get more probability maps of the object, which can provide more decisions for the final Bayesian update. For the dimly lit scene in the fourth scene of Fig. 6, our system shows a better segmentation effect on the approximate object than SemanticFusion.

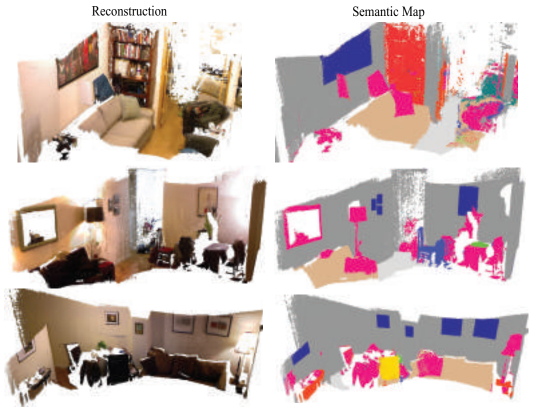

Fig. 7. The dense semantic map of our fusion algorithm on the NYUv2 dataset.

To fully verify the segmentation performance of our system, several more complex scenarios in the NYUv2 dataset are chosen for experiment. The experimental results are shown in Fig. 7.

In Fig. 7, the left are the reconstruction of the indoor scenes obtained by the traditional VGG and the ElasticFusion algorithm. The right are the semantic maps obtained by our DS-Net and ElasticFusion. Since ElasticFusion provides a precise camera trajectory and a more real-time reconstruction effect, continuous frames in the visual odometry and loop detection can establish a long-term correspondence between the landmark object and the map during the RGB-D camera movement. Our fusion algorithm can obtain a precise semantic segmentation image for the scene reconstructed by the ElasticFusion algorithm. Due to the accurate probability map provided by the DS-Net network, large objects in the scene, such as sofas, walls, floors, and our system have not been miss-segmented. For the small objects that appear continuously in the scene, they can be reconstructed more completely in our semantic map. In addition, our method realized the real-time construction of the indoor large-scale semantic map on the simulation dataset on the laptop, which proves that our system has better real-time performance.

To fully verify the real-time and effectiveness of our method, the experiment is carried out on our laboratory scenes with different distances from the sensor, the comparative results of our fusion algorithm and SemanticFusion are shown in Fig. 8.

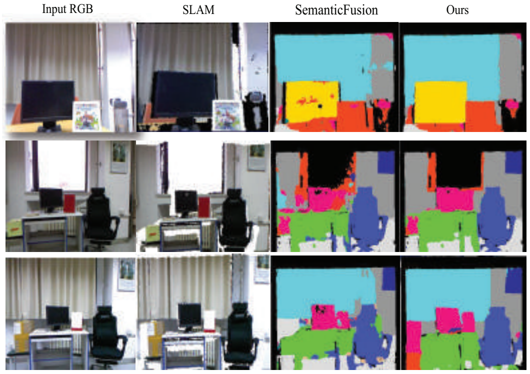

Fig. 8. Experients of the laboratory.

The first collumn in Fig. 8 are the input RGB images, the second image are the dense SLAM maps constructed by ElasticFusion algorithm, the third image are the semantic segmentation maps obtained using SemanticFusion, and the last image are the semantic segmentation maps obtained by our fusion algorithm. From the figure, we can see that different objects can be represented clearly in the segmented semantic map. As shown in the first scene, our method shows a better segmentation effect for the intersection parts of the curtain and the wall, and the display and the photo frame. In the second scene, our algorithm also shows superior results for tables and chairs with complex shapes. In the third scene, although the measured object is far from the sensor, we can see that our system not only has a better effect on large objects, such as windows, tables, chairs, but also has a better segmentation effect for the objects, such as the photo frame on the table and pictures hanging on the furthest wall. Due to the sufficient extraction and feature reuse of object information in our network, continuous frames in visual odometry and loop detection can also establish a long-term correspondence between landmark objects and dense maps, and so our system has better robustness and better performance for feature extraction and edge detection of complex objects.

V. Conclusions

In this paper, we propose an efficient semantic segmentation network, which realizes data fusion through Bayesian update by assigning a class probability to each surfel in the SLAM map. The experiments demonstrated that our method can effectively reduce the error of dense map construction and obtain the better effect of semantic segmentation. However, the reconstructed scene is not perfect for the map also contains many unrecognizable areas, which will result in a bad segmentation effect in the later semantic segmentation. Therefore, in the next work we will study the optimization problem of the ElasticFusion algorithm.

References

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

[2] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. ”SemanticFusion: Dense 3d semantic mapping with convolutional neural networks”. In: IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May-2 June 2017, pp.4628-4635. IEEE.

[3] Yang S, Huang Y, Scherer S. ”Semantic 3D occupancy mapping through efficient high order CRFs”. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, CANADA, 24-28 September 2017, pp. 590-597. IEEE.

[4] Hermans A, Floros G, Leibe B. ”Dense 3D semantic mapping of indoor scenes from RGB-D images”. In: IEEE International Conference on Robotics & Automation (ICRA), Hong Kong, CHINA, 31 May-7 June 2014, pp.2631-2638. IEEE.

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

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

[7] Dai A, Matthias Nie ner, Michael Zollh fer, et al. ”BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface re-integration”. ACM Transactions on Graphics, 2017, 36(4):1.

[8] Whelan T, Salas-Moreno R F, Glocker B, et al. ”ElasticFusion”. International Journal of Robotics Research, 2016, 35(14):1697-1716.

[9] Huang Gao, Liu Zhuang, van der Maaten, et al. ”Densely Connected Convolutional Networks”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 21-26 July 2017, pp.2261-2269. IEEE.

[10] H. Noh, S. Hong, and B. Han. ”Learning deconvolution network for semantic segmentation”. In: IEEE International Conference on Computer Vision (ICCV), Santiago, CHILE, 11-18 December 2015, pp.1520-1528. IEEE.

[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. ”Indoor segmentation and support inference from rgbd images”. In: European Conference on Computer Vision (ECCV), Florence, ITALY, 07-13 October 2012, pp.746-760.

[12] Bowman S L, Atanasov N, Daniilidis K, et al. ”Probabilistic data association for semantic SLAM”. In: IEEE International Conference on Robotics & Automation (ICRA), Singapore, 29 May-2 June 2017, pp.1722-1729. IEEE.

[13] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison. ”Slam++: Simultaneous localisation and mapping at the level of objects”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, 23-28 June 2013, pp.1352-1359. IEEE.

[14] Nakajima Y, Tateno K, Tombari F, et al. ”Fast and Accurate Semantic Mapping through Geometric-based Incremental Segmentation”. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, SPAIN, 01-05 October 2018, pp.385-392. IEEE.

[15] J. Long, E. Shelhamer, and T. Darrell. ”Fully convolutional networks for semantic segmentation”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), BOSTON, USA, 07-12 June 2015, pp.3431-3440. IEEE.

[16] Ivan Krešo, Josip Krapac, Siniša Šegvić. ”Efficient Ladder-style DenseNets for Semantic Segmentation of Large Images”. arXiv preprint arXiv:1905.05661, 2019.