3D Semantic Map Construction System Based on Visual SLAM and CNNs

1st Lei Lai 2nd Xinyi Yu 3rd Xuecheng Qian 4th Linlin Ou
College of Information Engineering, Zhejiang University of Technology
Hang Zhou, Peoples Republic of China
Hangzhou, China
joshualailei@163.com yuxy@zjut.edu.cn 306179250@qq.com linlinou@zjut.edu.cn

***Abstract*—Traditional approaches to simultaneous localization and mapping (SLAM) are unable to extract semantic information from the scene or meet the high-level task requirements of robots, and the efficiency of 3D map construction is low. To solve this problem, a 3D semantic map construction system is proposed in the paper to build a 3D semantic map. Firstly, the current position of the camera is estimated and optimized based on ORB-SLAM algorithm, and the globally consistent trajectory and pose are obtained. Then the semantic segmentation network is designed to predict semantic category of every pixel. The 3D semantic point cloud information is generated by combing the semantic information and the object point cloud. The global consistent camera pose estimated by visual SLAM algorithm is integrated with the semantic point cloud information to generate a 3D semantic map. Finally, we use an Octomap which is applied to navigation projects for map storage to reduce the storage capacity of the map. The experimental result verifies the accuracy and efficiency of this method.**

***Index Terms*—visual slam, semantic segmentation, semantic fusion, 3D semantic map**

I. Introduction

In recent years, with the development of computer technology and deep learning, Simultaneous Localization And Mapping (SLAM) based on vision has become an important research direction in the field of robots [1], [2]. SLAM utilizes the original information of the sensor to model the environment without any prior information of the environment and estimates the posture and movement of the robot at the same time. The robot position and map information generated by the traditional visual SLAM methods are only the connection of some dense or sparse geometric elements in the space. These spatial points provide relatively accurate position information, but there is no semantic attribute difference between them. However, to accomplish more complicated tasks, the robot needs to understand the surrounding objects and perceive the high-level semantic information of the environment. If the owner issues a command "fetch me a cup from the table", the robot needs to understand the semantic concepts of the table, the cup and their corresponding locations in the indoor environment. The semantic map contains the spatial geometry relationship in the environment. It can identify the independent individuals in the environment, and obtain the semantic information such as position, posture, and object category. Furthermore, the semantic map can improve the intelligence of human-computer interaction, and help the robot complete complex tasks.

SLAM algorithm is the foundation of semantic map construction. There are two main methods of visual SLAM: direct method and feature-based method. Engel proposed a direct monocular SLAM system LSD-SLAM [3], the transformation between frames is calculated by minimizing the grayscale difference of pixels. In feature-based methods, feature points in each image are extracted to match two images by minimizing feature reprojection errors. Klein [4] presented a SLAM method based on feature points called PTAM. It uses the Bundle Adjustment (BA) to complete the real-time SLAM system. Following the PTAM, ORB-SLAM [5] is a SLAM system based on ORB feature points, which can calculate the trajectory of camera in real-time and generate sparse 3D reconstruction results of the scene. Tracking, local mapping and relocation based on ORB feature points are performed and loop closing detection is used to eliminate accumulated errors.

However, it is not enough to obtain the position of camera for some complex tasks. As a main component of the semantic map, semantic segmentation has developed rapidly in recent years. Long [6] proposed a fully convolutional neural network (FCN) that allows to generate pixel-level prediction graphs and perform training in an end-to-end manner. At present, most advanced scene analysis frameworks are based on FCN. The deep convolutional neural network (CNN) method improves the ability to understand dynamic objects, but it still faces challenges in the case of diverse scenes and unlimited vocabulary. Zhao [7] presented PSPNet which solves the main problem of the FCN-based model, namely the lack of suitable strategies to utilize the global scene category clues. A pyramid pooling module is adopted to downsample the feature maps to different resolutions and then aggregate them to obtain better global context information.

With the rapid development of visual SLAM and deep learning, it is more possible to construct a 3D semantic map for localization, navigation, and automatic driving. Sunderhauf [8] proposed a CNN-based semantic scene recognition method that achieved scene classification by fusing 2D lidar and camera data. However, only the category attributes of the point cloud can be obtained, while objects in the category

978-1-7281-5414-5/20/$31.00 ©2020 IEEE

cannot be distinguished. Vineet [9] proposed a method for online dense reconstruction of semantic labeling. However, they only focused on building maps and segmentation, while semantic information was not utilized for map construction. Kundu [10] established a global CRF model with voxel by combining the reconstruction and the semantic annotation of a 3D map, but the model is not easy to produce results. Bao [11] integrated camera parameters, object geometry information, and object category information into a Structure From Motion (SFM) problem to form a detailed but computationally intensive optimization problem. SLAM ++ [12] focuses on the map construction of indoor scenes at the semantically defined object level. Models of known objects are identified and inserted by using RGB-D sensor information, and then these object models are utilized as landmarks for tracking and mapping. J. McCorma proposed a dense 3D semantic mapping method using convolutional neural networks, called SemanticFusion [13]. It utilizes ElasticFusion [14] as SLAM backend and fuses semantics information on each surfel to produce a surfel based 3D semantic map. But the point cloud map is too large and difficult to for navigation.

In this paper, we design a 3D semantic map construction system based on visual SLAM and convolutional neural networks, which can make the robot achieve more intelligent navigation tasks. Firstly, we use the RGB-D camera as a sensor, and adopt the ORB-SLAM2 algorithm to estimate and optimize the current position of camera. Then the semantic segmentation algorithm is used to perceive the semantic information of the environment and the category and position of the object is detected. Then we combine the object point cloud with the semantic information to generate 3D semantic point cloud information. Finally, the globally consistent camera poses estimated by the visual SLAM algorithm are used to perform voxel-based semantic fusion to generate a 3D semantic map including geometric information and semantic information. The high efficiency of the system is verified by experiments. In this paper, we design a semantic segmentation model to perform pixel-level semantic annotation. Meanwhile, we provide an Octomap for storing maps, which can reduce the storage space of the map and can be employed for high-level tasks such as navigation.

II. METHOD

The overall framework of the 3D semantic map construction system based on visual SLAM and convolutional neural network is shown in Fig.1. The entire system receives color and depth images and outputs 3D semantic maps. First, the color images and depth images of RGB-D camera are input into two different processes. One is the visual SLAM process, which can get the camera pose at every moment by matching the feature points between adjacent frames. The other is the semantic point cloud generation process. It performs semantic segmentation on the input color image and the point cloud is generated according to the input depth image and camera intrinsic matrix. The semantic color and the original color information are then added to the generated point cloud. Finally, voxel-based semantic fusion was performed to generate an octree map based on the produced point cloud. Meanwhile, the 3D semantic map containing geometric information and semantic information is constructed and sent to a program (rviz) for visual display in real-time.

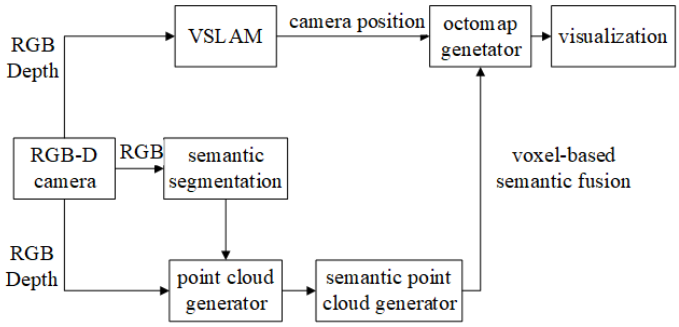


Fig. 1. The framework of proposed 3D semantic map construction system.

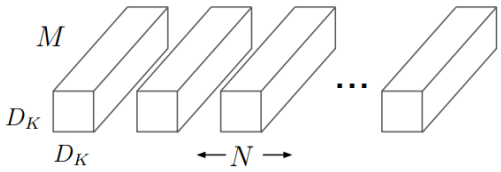
A. SLAM Mapping

The proposed semantic mapping system requires knowledge of the current camera pose, which can be provided by a SLAM system. In this paper, we choose ORB-SLAM which uses both RGB and depth images for sparse tracking as our SLAM backend. ORB feature points in the current frame image are extracted to match the feature points of the previous frame image. According to these two sets of matched feature points, an optimization problem is established to calculate the pose of the current frame. As a consequence, the computed camera transform is broadcasted in the system and received by the Octomap generation process.

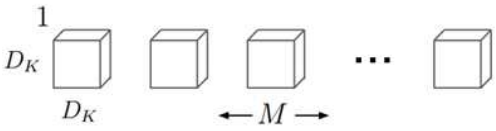
B. Object Detection for Semantic Mapping

In this paper, We utilize PSPNet as our CNN model. First, the image is input to a convolutional neural network which is used as a feature extraction network to extract image features. The input image gets the feature image of the original image size 1/8 through the feature extraction network. The feature image is sent to the pyramid pooling module. The pyramid pooling is divided into 4 different scales. After pooling, feature maps of different sizes can be obtained. For each pyramid level feature map, we perform 1x1 convolutional dimensionality reduction operation, and then directly upsample the low-dimensional feature map to obtain the original image size. Finally, the feature maps of different layers are merged with the original feature map to form the feature map and final convolutions are performed to produce the class score map.

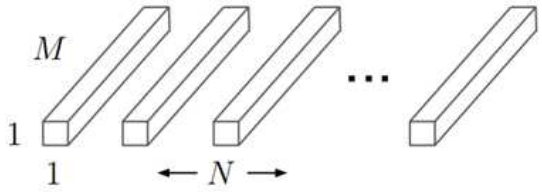
In the semantic map construction system of this paper, the semantic segmentation module provides us with original semantic information. To ensure the real-time performance of the entire system, especially the visual SLAM module, the computation of the semantic segmentation module needs to be reduced as much as possible. Therefore, we compress the semantic segmentation network structure and reduce the model parameters in this paper.



(a) Standard Convolution Filters



(b) Depthwise Convolution Filters



(c) 1x1 Convolution Filters called Pointwise Convolution

Fig. 2. Depthwise Convolution Filters.

Sifre [15] proposed a depthwise separable convolution to reduce the parameters of the model. Depthwise separable convolution decomposes standard convolution to obtain depthwise convolution and 1x1 pointwise convolution. The decomposition process is shown in Fig.2. Assume that the input feature map dimension is $D_F \times D_F \times M$, and the convolution kernel size is $D_k \times D_k$. Where D_F and D_G are respectively the size of the input and output images, M and N are the number of input and output channels respectively, and D_k represents the size of the convolution kernel. For a standard convolutional filter, the calculation formula of the convolution operation is ${D_k}^2 \times M \times N \times {D_F}^2$

After applying depthwise separable convolution, the standard convolution kernel is decomposed into M convolution kernels of $D_k \times D_k \times 1$ and N convolution kernels of $1 \times 1 \times M$. The computational cost is ${D_k}^2 \times M \times {D_F}^2 + M \times {D_F}^2$

As a result, the ratio of the two calculations is:

$$\frac{{D_k}^2 \times M \times N \times {D_F}^2}{{D_k}^2 \times M \times {D_F}^2 + M \times {D_F}^2} = \frac{1}{N} + \frac{1}{D_k^2} \tag{1}$$

This shows that the depthwise separable convolution can greatly reduce the amount of model calculation.

In this section, we modify the feature extraction layer of PSPNet-50 by replacing Resnet-50 with MobileNetV2 [16]. MobileNet utilizes a deepwise separable convolution method to reduce the size of the model and achieve good results in image segmentation tasks. The network structure is shown in Fig.3. First input the image to the MobileNetV2 network to speed up the feature map extraction, then connect the pyramid pooling structure, and finally output through the convolution layer.

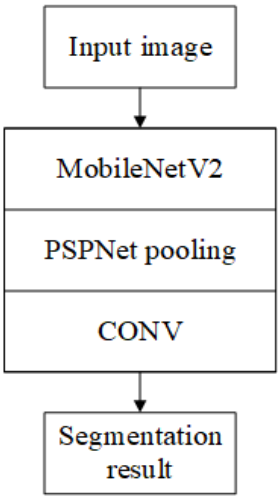


Fig. 3. Network structure of this paper.

C. semantic mapping

After receiving the image information by RGB-D camera, the proposed system performs point cloud generation, semantic segmentation, and camera pose estimation. After the semantic segmentation is performed, the recognized target semantic color is added to the point cloud to generate 3D semantic point cloud information. The 3D semantic point cloud information contains the 3D coordinates of the object in the world coordinate system and semantic information. We use it to build a semantic map.

To reduce the storage capacity of the map and allow the map to be utilized in navigation projects, we choose Octomap [17] for map storage. Octomap is a probabilistic map framework based on an octo-tree structure, which can compress point clouds and save storage space. The structure of octo-tree is shown in Fig.4. The 3D space is continuously cut based on the octo-tree structure until it becomes the smallest square. In an octo-tree, a node stores information about whether it is occupied. When all child nodes of a block are occupied or unoccupied, there is no need to expand the node. A floating-point number between 0 and 1 is used for each small square to express the probability that it is occupied.

During the observation of the environment, the nodes need to be updated probabilistically due to the influence of noise or dynamic objects. The formula is given as follows:

$$P(n|z_{1:T}) = \left(1 + \frac{1 - P(n|z_T)}{P(n|z_T)} \frac{1 - P(n|z_{1:T-1})}{P(n|z_{1:T-1})} \frac{P(n)}{1 + P(n)}\right)^{-1} \tag{2}$$

where n represents the leaf node, z_T represents the measured value at time T, and $P(n|z_{1:T})$ represents the occupancy probability of the leaf node. In order to ensure that the probability is between 0 and 1, the above formula is written

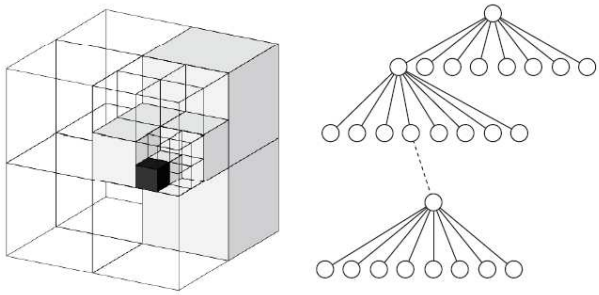


Fig. 4. Structure of octo-tree.

(a) (b) (c)

Fig. 5. This figure shows the semantic segmentation results of the two methods (a) color image. (b) PSPNet-50. (c) MobileNetV2-PSPNet.

in the form of logarithm of probability and brought into the formula:

$$y = \log it(x) = \log(\frac{x}{1-x}) \tag{3}$$

Converting (2) yields:

$$L(n|z_{1:t+1}) = L(n|z_{1:t-1}) + L(n|z_t) \tag{4}$$

As a result, the current observation data is continuously fused to achieve node updates, and finally the Octomap is created. To store the 3D semantic point cloud information in Octomap, the specific process is as follows: first, calculate the free node between the camera pose and the target object, then insert the free point and the 3D semantic space point into the map, and update the probability of the corresponding voxel value. At the same time, the basic storage unit is represented by voxels, and each voxel simultaneously stores the probability of occupation and the scores of different object categories corresponding to the voxel.

In the process of updating the octree map, semantic segmentation of the same voxel may cause inconsistent labels due to the uncertainty of the sensor and the environment. When fusing voxels between different frames, inconsistent labels will be generated conflict. To solve this problem, this paper adopts the Bayesian fusion method to perform voxel-based semantic label fusion. Bayesian fusion is a widely used method in multi-view semantic fusion. The formula is given as follows:

$$p(y|z^i) = \frac{p(z_i|y, z^{i-1})p(y, z^{i-1})}{p(z_i|z^{i-1})} \tag{5}$$

where y refers the semantic label information of the current voxel, and z_i refers the measurement value of the voxel in ith frame.

$$p(y|z^i) = \eta_i p(z_i|y, z^{i-1})p(y|z^{i-1}) \tag{6}$$

We use Bayesian fusion to take a voxel as a unit to semantically label a single frame, and then normalize it to obtain an effective probability distribution. This incremental fusion of semantic probability information enables the system to utilize all existing frames to update and optimize the semantic labels of 3D points in real-time, improving the accuracy of semantic annotation.

III. EXPERMINE

The composition of the platform in the experiment mainly includes one Xtion camera, one computer with Ubuntu system installed, Intel Core i7-7800K 3.50Ghz, and Nvidia TITAN Xp.

A. Network training

To verify the efficiency of the proposed method, the improved MobileNetV2-PSPNet was tested on the ade20K [18] dataset. In the process of model training, 32 batch size was used, and the initial learning rate was 0.001. The following three indicators were used to evaluate the model in this paper: (1) pixel accuracy; (2) IoU score, that is, the average score of the joint intersection of each category,$IoU = \frac{TP}{TP+FP+FN}$ where TP, FP, and FN are pixels whose predicted results are real, false, and false negative, respectively. (3) the time it takes to train an image. (4) the size of the network model after training. The model training results are shown in TAB.1. The result of semantic segmentation is shown in Fig.5.

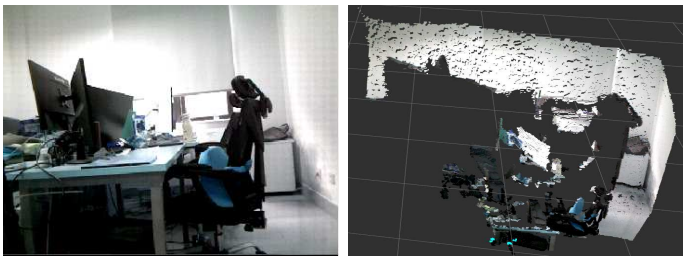
TABLE I
COMPARISON RESULTS OF THIS PAPER AND PSPNET-50.

Model name	Pixel Acc.(%)	Mean IoU(%)	time(ms)	size(M)
$Ours$	79.31	37.23	67	59
$PSPNet-50$	79.73	41.26	120	203

According to the above comparison and analysis, compared with PSPNet-50, the MobileNetV2-PSPNet network designed in this paper can reduce the storage size of the network model to about 1/4 and increase the computing speed by about one time while maintaining a small difference in accuracy. Overall, it can meet the requirements of accuracy and efficiency of the system.

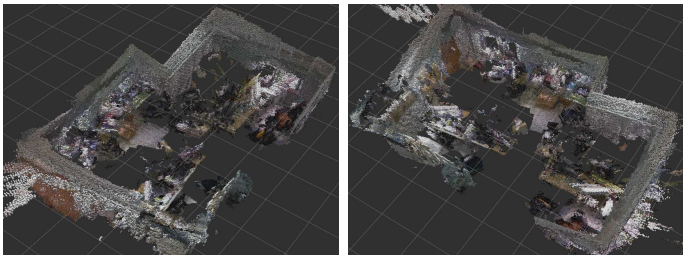
B. 3D reconstruction

We use the calibrated xiton camera collecting color and depth information of the scene to generate a point cloud image. An example of point cloud generation is shown in Fig.6. Then the camera pose information obtained by ORB-SLAM algorithm and point cloud information are combined to obtain a 3D point cloud map. The 3D point cloud map is shown in Fig.7.



(a) (b)

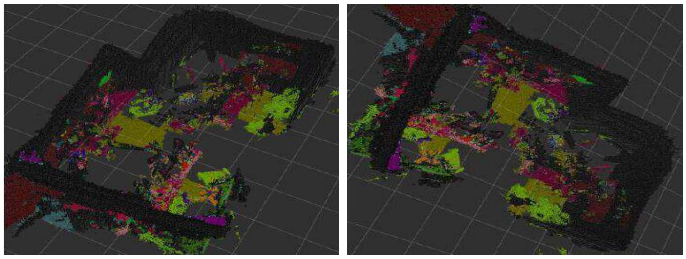
Fig. 6. This figure shows the generation of point cloud. (a) lab scene (b) point cloud.



(a) (b)

Fig. 7. Reconstructed scene with RGB color.

(a) (b)



(c) (d)

Fig. 8. (a) the color image of the current camera.(b) the 2D recognition results obtained through the semantic segmentation model.(c)(d) the scene semantic map finally obtained.

C. Semantic map

We test our semantic mapping system in a laboratory environment. We start the calibrated xiton camera and run the system. During the operation of the system, the poses of the camera are estimated by the visual SLAM module in real-time. Meanwhile, the semantic segmentation module acquires the position and category attributes of objects, and finally an octo-tree semantic map is constructed. After moving and rotating the handheld camera for a while in the lab, a 3D map of the scene is constructed, and the detection results are shown in Figs 8 and 9. It can be seen that this system can accurately identify indoor objects.

IV. Conclusion

We propose an efficient semantic segmentation model, which provides a guarantee for the real-time performance of the entire system. We use an Octomap which is applied to navigation projects for map storage to reduce the storage capacity of the map. We use the Bayesian probability method to optimize the fusion of the semantic probability labels of each voxel, thereby increasing the accuracy of 3D semantic annotation.In fact,there are still some challenging tasks. The semantic SLAM system in this paper is based on the assumption of a static environment, we will further study how to identify the dynamic objects and exclude them. In addition, RGB-D camera was only used in our paper, and the positioning accuracy will be more dependent on the trajectory of camera. In the future, we will further study the multi-sensor visual SLAM fusion method to improve the accuracy of mapping.

V. Acknowledgements

This paper was supported by National Key Research and Development Plan Intelligent Robot Key Project (2018YFB1308400).

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] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendn-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[3] J. Engel, T. Schps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," 2014.

[4] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *IEEE and Acm International Symposium on Mixed and Augmented Reality*, 2007.

[5] R. Mur-Artal, J. Montiel, and J. D. Tardos, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2016.

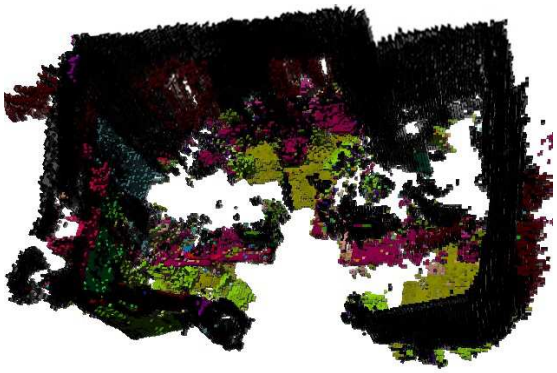
[8] N. Snderhauf, F. Dayoub, S. Mcmahon, B. Talbot, R. Schulz, P. Corke, G. Wyeth, B. Upcroft, and M. Milford, "Place categorization and semantic mapping on a mobile robot," 2015.

[9] V. Vineet, O. Miksik, M. Lidegaard, M. Niener, S. Golodetz, V. A. Prisacariu, O. Khler, D. W. Murray, S. Izadi, and P. Prez, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[10] A. Kundu, L. Yin, F. Dellaert, F. Li, and J. M. Rehg, "Joint semantic segmentation and 3d reconstruction from monocular video," 2014.

[11] Y. Z. Bao and S. Savarese, "Semantic structure from motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

(a) Resolution 0.02 m



(b) Resolution 0.04 m

(c) Resolution 0.08 m

Fig. 9. Semantic map at different resolutions.

[12] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[13] J. Mccormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," 2016.

[14] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "Elasticfusion: real-time dense slam and light source estimation," *International Journal of Robotics Research*, vol. 35, no. 14, 2016.

[15] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *Computer ence*, vol. 3559, pp. 501–515, 2014.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[17] 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.

[18] B. Zhou, Z. Hang, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.